Multi-Model Assessment of Regional Surface Temperature Trends: CMIP3 and CMIP5 20th Century Simulations

Thomas R. Knutson, Fanrong Zeng, and Andrew T. Wittenberg

Geophysical Fluid Dynamics Laboratory/NOAA, Princeton, NJ 08542

Journal of Climate (in press)

Version: Matches June 2013 Page Proof version, created Sept.. 2013

Email contact: Tom.Knutson@noaa.gov

Abstract. Regional surface temperature trends from the CMIP3 and CMIP5 20th century runs are compared with observations -- at spatial scales ranging from global averages to individual grid points -- using simulated intrinsic climate variability from pre-industrial control runs to assess whether observed trends are detectable and/or consistent with the models' historical run trends. The CMIP5 models are also used to detect anthropogenic components of the observed trends, by assessing alternative hypotheses based on scenarios driven with either anthropogenic plus natural forcings combined, or with natural forcings only. Modeled variability is assessed via inspection of control run time series, standard deviation maps, spectral analyses, and low-frequency variance consistency tests. The models are found to provide plausible representations of internal climate variability, though there is room for improvement. The influence of observational uncertainty on the trends is assessed, and found to be generally small compared to intrinsic climate variability.

Observed temperature trends over 1901-2010 are found to contain detectable anthropogenic warming components over a large fraction (about 80%) of the analyzed global area. In about 70% of the analyzed area, the modeled warming is consistent with the observed trends; in about 10% it is significantly greater than simulated. Regions without detectable warming include the high latitude North Atlantic, the eastern U.S., and parts of the eastern and northern Pacific. For 1981-2010, the observed warming trends over only about 30% of the globe are found to contain a detectable anthropogenic warming; this includes a number of regions within about 40-45 degrees of the equator, particularly in the Indian Ocean, western Pacific, South Asia, and tropical Atlantic.

1. Introduction

Are historical simulations of surface temperature trends, obtained using climate models with the best available estimates of past climate forcings, consistent with observations? Where on the globe can observed temperature trends be attributed to anthropogenic forcing? These questions can be examined using a substantial number of different climate models and using different analysis methods. Here we attempt to incorporate information from a relatively large sample of

climate models, from the Coupled Model Intercomparison Project 3 (CMIP3; Meehl et al. 2007) and CMIP5 (Taylor et al. 2012), using various multi-model combination techniques. The general approach is to compare the modeled and observed trends, in terms of both magnitude and pattern, by considering trends at each grid point in the observational grid, as well as trends over broader-scale regions.

The term "*detectable climate trend*" used here refers to a trend in the observations that is inconsistent with (i.e., outside of the 5th to 95th percentile range of) simulated trends, either from control runs (the internal or intrinsic climate variability background) or from a sample of natural-forcing response and control run variability combined (the natural climate variability background). (Control runs are long runs with pre-industrial forcings that do not change from year to year.) We interpret a trend in observations as "*attributable (at least in part) to anthropogenic forcing"* if it is both inconsistent with simulated natural climate variability (detectable) and consistent with the All-Forcing runs that contain both anthropogenic forcing agents (e.g., changes in greenhouse gases and aerosols) and natural forcings (e.g., changes in solar insolation or volcanic aerosol loading). If an observed trend is detectable but inconsistent with All-Forcing runs because it is larger than the simulated distribution of trends, we still interpret the observed trend as attributable, at least in part, to anthropogenic forcing. While a number of CMIP5 models have Natural-Forcing-Only runs available on-line, for the CMIP3 models, relatively few such runs are available. Therefore, for CMIP3, we adopt a simpler approach of assessing whether observed trends are consistent with All-Forcing runs, but inconsistent with internal variability alone. The simpler approach does not allow us to draw conclusions about whether an observed trend is attributable to anthropogenic forcing or not.

The modeled internal climate variability from long control runs is used to determine whether observed and simulated trends are consistent or inconsistent. In other words, we assess whether observed and simulated forced trends are more extreme than those that might be expected from random sampling of internal climate variability. This approach has been applied to earlier models in a number of studies, beginning with the analyses of Stouffer et al. (1994; 2000). Similarly, we use the available ensemble of simulated forced trends to assess whether observed trends are compatible with the forcing-and-response hypotheses embodied by those forced simulations.

Formal detection/attribution techniques often use a model-generated pattern from a single or set of climate forcing experiments, and then regress this pattern against the observations to compute a scaling amplitude (e.g., Hegerl et al. 1996; Hasselmann 1997; Allen and Tett 1999; Allen and Stott 2003) . If the scaling is significantly different from zero, the forced signal is detected. If the scaling does not significantly differ from unity, then the amplitude of the signal agrees with observations, or is at least close enough to agree within an expected range based on internal climate variability. Optimal detection techniques also filter the data during the analysis such that the chance of detecting a specified signal, or "fingerprint", is enhanced if the signal is present in the data. An alternative approach that is less focused on model-defined patterns has been proposed by Schneider and Held (2001). In contrast to the optimal detection/attribution methods, we compare both the amplitude and pattern simulated by the models directly with the observations, without rescaling of patterns or application of optimization filtering. Our analysis is thus a consistency test for both the amplitude and pattern of the observed versus simulated trends, building on earlier work along these lines by Knutson et al. 1999; Karoly and Wu 2005;

Knutson et al. 2006; and Wu and Karoly 2007 to test for detectable anthropogenic contributions. Other variants and enhancements to this general type of analysis have recently been presented by Sakaguchi et al. (2012). More discussion of various detection and attribution methods and their use in general is contained in Hegerl et al. 2009.

In this report, the models, methods, and observed data are described in Section 2. We examine the model control runs and their variability in Section 3. Global-mean time series from the 20C3M (approximately 1860-2010) historical runs are examined in Section 4. Section 5 contains consistency tests for observed vs simulated trends, as discussed above, for temperatures averaged over various defined regions of the globe. Maps based on results of consistency tests at the grid point scale are presented in Section 6. A brief description of online supplemental material is given in Section 7, and the discussion and conclusions are given in Section 8.

2. Model and Observed Data Sources

a. Observed data

The observed surface temperature dataset used in this study is the Hadley Centre/Met Office and Climatic Research Unit/University of East Anglia combined land and sea surface temperature version 4 (HadCRUT4; Morice et al. 2012) which is available as a set of anomalies relative to the period 1961-1990. The dataset contains some notable revisions, particularly to SSTs (HadSST3; Kennedy et al. 2011), relative to previous versions, so it important to retest earlier conclusions regarding climate trends using the revised data. The dataset also contains uncertainty information, in the form of 100 ensemble members sampling the estimated observational uncertainty. Some of our tests examine the sensitivity of trend results to this observational uncertainty.

To form a combined product of SST and land surface air temperature, Morice et al. (2012) adopt the following procedure. If both land data and SST data are available in a particular grid box, they are weighted according to the fraction of the grid box that is covered by land or ocean, respectively. A minimum of 25% coverage is assumed, even if the fraction of the grid box covered by land is less than 25%. In our study, we use this same general procedure, adapted to a model's land-sea mask, to combine SST and land surface air temperature data sets from each model that we analyze.

b. CMIP3 and CMIP5 models

Figure 1 displays the complete collection of control runs from both CMIP3 and CMIP5 used in our analysis. The data were downloaded from the CMIP3 (www-pcmdi.gov/ipcc/about_ipcc.php) and CMIP5 (cmip-pcmdi.llnl.gov/cmip5) model archives. We regridded (averaged) the model data from the 20C3M historical runs and control runs onto the observational grid. In cases where we needed to use combined model land surface air temperature and SST data to compare with observations, we used a procedure resembling that

used for the observations, but based on the model's own land-sea mask. For example, if any land is present in a grid box, a minimum of 25% land coverage is assumed, even if the fraction of the grid box covered by land is less than 25%. Our general approach in this study is to attempt to mimic observations with the models, in terms of data coverage over time. To mimic the space-time history of data gaps in the observations, we masked out (withheld from the analysis) model data at times and locations where data were labeled missing in the observations. Finally, we computed the model's climatology over the same years as for observations (1961-1990) and then created anomalies from this climatology. For example, this same procedure was used for 150-yr samples from the model control runs for analyses where we wanted to ensure that the control runs had missing data characteristics that were similar to those of the observed data. The sea surface temperature data from CMIP3 is based on each ocean model's top layer; for CMIP5 it is based on the skin temperature due to difficulties in using native ocean model grid data from the archive.

The historical forcings for the CMIP3 20C3M historical forcing runs are summarized in Rind et al. (2009; Table 3.6). An important distinction among the models is the treatment of volcanic forcing. Ten of the 24 CMIP3 models we examined include volcanic forcing, while 14 do not. However, as discussed further below, for most of our assessments, we used a maximum of 19 of the 24 CMIP3 models of which eight included volcanic forcing while 11 models (identified by "*" after model name in Fig. 1 a,b) did not. We refer to these sets of models as the eight "Volcanic" and 11 "Non-Volcanic" CMIP3 model subsets, respectively. All 23 of the CMIP5 models included in this study incorporated volcanic forcing in their 20C3M runs. However, only seven of the 23 CMIP5 models had Natural-Forcing-Only runs that extended to 2010 (see Fig. 1). These Natural-Forcing runs extending to 2010 were necessary for some of our detection and attribution analyses concerning anthropogenic forcing, and those seven models form the CMIP5 seven-model subset referred to in subsequent sections.

3. Model Control Run Analysis

a. Global mean time series

The global-mean surface air temperature series from the CMIP3 and CMIP5 model control runs are shown in Fig. 1. Data are displayed with arbitrary vertical offsets for visual clarity. The figure also shows the observed surface temperature anomalies from HadCRUT4. The curves labeled "Residual" were obtained by subtracting the multi-model mean of the historical (volcanic) forcing runs (either CMIP3 or CMIP5) from the full observed time series. These observed residual series thus contain estimates of the internal variability of the climate system as derived from the observations in combination with the climate models' response to estimated historical forcing. In section 3b we will further refine this estimate of observed internal variability.

The model control runs exhibit long-term drifts. The magnitudes of these drifts tend to be larger in the CMIP3 control runs (Fig. 1a,b) than in the CMIP5 control runs (Fig. 1 c,d), although there are exceptions. We assume that these drifts are due to the models not being in equilibrium with the control run forcing, and we remove the drifts by a linear trend analysis (depicted by the orange straight lines in Fig. 1). In some CMIP3 cases the drift initially proceeds at one rate, but

then the trend becomes smaller for the remainder of the run. We approximate the drift in these cases by two separate linear trend segments, which are identified in the figure by the short vertical orange line segments. These long-term drift trends are removed to produce the "drift-corrected" series. The procedure for removing the trends involves calculating and removing the linear trends (over the time periods shown in Fig. 1) at each model grid point separately. The orange trend lines shown in Fig. 1 depict also the starting and ending years for the trends used for each model.

Five of the 24 CMIP3 models, identified by "(-)" in Fig. 1, were not used, or practically not used, beyond Fig. 1 in our analysis. For instance, the IAP_fgoals1.0.g model has a strong discontinuity near year 200 of the control run. We judge this as likely an artifact due to some problem with the model simulation, and we therefore chose to exclude this model from further analysis. The Miroc_3.2_hires and INGV_echam4 model control runs are so short in length that they are essentially unused in our analysis, since we require the control run record to be at least three times as long as a trend that is being assessed. For two other models, we were not able to successfully obtain sea surface temperature information from the CMIP3 archive, and so these were excluded from further analysis.

While some of the trends in the CMIP3 and CMIP5 control runs (Fig. 1) approach the observed ~150 yr trend in terms of general magnitude, these few cases are associated with either the long-term drifts discussed above or with a few spurious discontinuity issues (e.g., IAP_fgoals1.0.g). Controlling for these apparent problems, none of the control runs in the CMIP3 or CMIP5 samples exhibit a centennial scale trend as large as the trend in the observations. On the other hand, the variability of observed residual series appears roughly similar in scale to that from several of the control runs. Three of the CMIP3 control runs illustrated in Fig. 1 (GISS_aom, GISS_model_e_h, and GISS_model_e_f) have much lower levels of global surface temperature variability than in the observed residual series. For some sensitivity tests on the multi-model assessments, we have excluded these three models to test for robustness.

b. Geographical distribution of variability

In this section, we describe a method for comparing the geographical distributions of observed variability with model control run variability. The geographical distribution of an adjusted standard deviation of low-pass-filtered (> 10 yr) surface temperature from observations (Obs. St. Dev.*) is shown in Fig. 2 (middle column: b, e, h). These observed estimates contain adjustments (described in detail below) that make them more suitable for comparison to the variability in the model control run. This is necessary because the variability within the model control runs is generated strictly internally within the models and does not contain contributions from external climate forcings. In contrast, observed temperature will contain some mixture of variability due to external climate forcing agents and internally generated processes in the climate system. The models' average standard deviation fields, based on the full available time series of surface air temperature from each control run, are shown in the left column (a, d, g). Prior to computing the individual model standard deviations, the long-term drift has been subtracted from each control run as discussed in Section 3a. The individual model standard deviations are then averaged for the three model sets to form the fields in (a, d, g). Difference

maps, computed as the models' average low-frequency standard deviations minus Obs. St. Dev*, are shown in the right column (c, f, i) of Fig. 2.

We now describe the process for computing the adjusted observed low-pass filtered standard deviation (Obs. St. Dev.*; Fig. 2 b, e, h). At each gridpoint, we low-pass filter the observations using a decadal filter with a half-power point at nine years. Rather than compare this variance directly to variance from a model control run, we first attempt to estimate how much of an amplification of variance there is in the observed estimate owing to the presence of forced variability (in addition to internal, unforced variability). We then correct or adjust for this amplification in two stages. For each of the three sets of models (CMIP3 eight-model set; CMIP5 23-model set, and CMIP5 seven-model subset), analyzed separately, we use the grand ensemble mean of the model All-Forcing runs (n=8, 23, or 7) as an estimate of the forced signal to remove from observations. This provides the "first level" adjustment for the observations, which is slightly different for each set of models. However, since the true forced response of a given model is only approximately known, given the limited number of ensemble members that are used to estimate this forced response, it follows that some residual forced variance will remain in the observed series after this initial adjustment. We try to estimate how much variance remains by using the same procedure that we used for observations, but applying it to each individual All-Forcing run ensemble member. That is, for a given model, we consider each of its All-Forcing ensemble members separately and remove the multi-model ensemble mean (as for observations) to derive an internal variability estimate. We average this estimate across all of that model's ensemble members to create an average standard deviation for that model, and then average across all models to create a multi-model ensemble internal variability standard deviation estimate. Next, we consider the model control runs, and compute the average standard deviations for a sample of 50 randomly drawn 110-yr time series from each control run, average those, and then average across all of the control runs to create an ensemble-average internal variability estimate from the control runs. For each of the 110-yr segments, the control model data is masked with the observed mask for the given grid point before being low-pass filtered. By comparing the internal variability estimate derived from the All-Forcing runs with that from the control runs for the same models, we derive the "second-level" adjustment. This average adjustment is then applied to the standard deviation from the "first-level" adjusted observations to obtain a new observed internal variability standard deviation estimate (Obs. St. Dev.*) that is more suitable to compare with the model control runs. Given this method (which includes two separate levels of adjustment) we can now more defensibly compare the model control run and observed low-frequency variability.

We stress that our variance-comparison procedure described above is only a very rough test of decadal variance consistency, and is not even attempted in data-poor regions such as the deep Southern Ocean. There are inherent limitations to our estimates because there is only so much observational data and only so many ensemble members supplied by the modeling centers. In terms of observational temporal coverage, in order for a comparison to be done between model and observations at a grid point, we require at least 50 points (out of 110) to be available in the 110-year annually resolved decadally filtered record. Forty percent temporal coverage is required for an annual mean to be considered valid, and the decadal filter does a modest degree of gap-infilling by computing a filtered value if at least four of seven annual values are available within a seven-year wide sliding window.

The adjusted standard deviation of low-pass filtered observations ("Obs. St. Dev.*") forms the basis of the observed estimates and difference maps in Figs. 2 and 3, and of the variance consistency tests that will be described later in this report.

The adjusted observed fields ("Obs. St. Dev.*") suggest that the strongest low-frequency internal surface temperature variability occurs over higher latitude land and oceanic regions of the Northern Hemisphere. The modeled fields also show these features, though they are somewhat stronger in the models than for the observed estimate. Thus, a feature that stands out in the modeled minus observed (Obs. St. Dev.*) standard deviation field (Fig. 2 c, f, i) is the tendency for model-simulated low-frequency internal variability to exceed the observed estimate in high-latitude oceanic and continental regions of the Northern Hemisphere. Another feature is a tendency for the modeled variability to be too small over much of the remaining ocean regions and Southern Hemisphere as far south as about 40^{o}S. Limited data coverage precludes an assessment of low-frequency variability over much of the Arctic Ocean, Antarctica, and the Southern Ocean south of 40^{o}S (gray regions on the maps).

The general features shown in the ensemble mean difference maps in Fig. 2 (c, f, i) are also present to some degree for many of the individual models (Fig. 3). We also list in Fig. 3 the spatial correlation coefficients between the individual model standard deviation fields (not shown) and the observed field (Obs. St. Dev.*). These spatial correlations typically vary from about 0.5 to 0.7 for the models shown, indicating a relatively good agreement between individual models and observations in the overall spatial structure of the variability. This gives us some confidence in the models' ability to simulate at least the broad-scale features of surface temperature low-frequency variability.

There are a number of caveats to the comparison presented here. For example, uncertainties remain in estimating the forced variability component from observations, which is used to create the observed residual, and thus there are uncertainties in the observed internal variability estimate used for comparison to the model control runs, as noted earlier. In addition, the available observational records are relatively short compared with many of the model control runs. As noted by Wittenberg (2009) and Vecchi and Wittenberg (2010), long-running control runs suggest that internally generated SST variability, at least in the ENSO region, can vary substantially between different 100-yr periods (approximately the length of record used here for observations), which again emphasizes the caution that must be placed on comparisons of modeled vs. observed internal variability based on records of relatively limited duration.

4. Global mean surface temperature: Historical forcing runs

a. Time series of global mean surface temperature

The global mean time series of surface temperature from the 20C3M historical runs are compared with observations (black curves) in Fig. 4 in a form similar to that presented by Hegerl et al. (2007). The historical dates of large volcanic eruptions are shown by vertical brown lines. An analysis of the model time series for the CMIP3 and CMIP5 All-Forcing experiments is presented in Figs. 4a-c, and for the available CMIP5 Natural-Forcing-Only experiments in Fig. 4d. The large shaded region on each plot shows the 5^{th} to 95^{th} percentile range of a single model

realization from the multi-model sample. The multi-model sample is formed by combining the distributions of each of the models, with each model having an equal probability weight in the multi-model distribution. The sub-distribution from each model is centered on that model's ensemble mean with the distribution about that mean based on the control run for that model. Thus the multi-model distribution incorporates the uncertainty due to differences between the model ensemble means (i.e., forcing and response-to-forcing uncertainties) and uncertainties due to internal variability for each model.

The analysis shows that for the All-Forcing runs (Fig. 4 a-c) most of the time the observed annual means lie within the 5th to 95th percentile range of single model realizations, implying that there is a consistency between the observed record and the multi-model ensemble of runs taken as a whole. However, the range for the CMIP5 Natural-Forcing-Only simulations (Fig. 4d) clearly separates from the observed time series after about 1960, indicating that Natural-Forcing-Only runs are inconsistent with observations, particularly for the late 20th century global warming.

The narrower shaded region between the two thick red lines (a-c) depicts the 5th to 95th percentile range of the multi-model ensemble mean. This is fairly narrow, indicating that the multi-model ensemble means of these particular sets of models are fairly well-constrained, with relatively small uncertainty. The ensemble means of the CMIP3 and CMIP5 volcanic models (Fig. 4 a,c) track the observations remarkably well although the apparent volcanically induced temporary dips are not in full agreement with the observed behavior for those periods. For example, in Fig. 4a, and 4c, the multi-model responses to the Pinatubo and Krakatau eruptions appear to be larger than in observations. These apparent discrepancies in the volcanic responses will require further analysis (see e.g. Stenchikov et al. 2009) and are not a focus of the present study. For example, one must carefully assess the role of internal climate variability in judging whether these differences are significant or not.

The combined volcanic and non-volcanic CMIP3 ensemble (Fig. 4 (b)) shows a substantially wider envelope of model behavior, as expected with the larger number of models and with the wider discrepancy in forcing among these models. Since the "Non-Volcanic" runs have a substantially less realistic representation of the forcing, we will generally emphasize the eight CMIP3 models with "Volcanic" runs in panel (a) in our remaining forced model assessments for the CMIP3 models in this study.

b. Spectra of global mean surface temperature

Figure 5 (a,b) shows the variance spectra of observed global mean temperature (black curves, with a shaded range for the 90% confidence intervals) and of the individual CMIP3 and CMIP5 "Volcanic forcing" historical runs (red curves) from Fig. 4 (a, c), using data from the years 1880-2010. The data were not detrended prior to computing the spectra. Before plotting, the raw spectra were smoothed using a non-overlapping sliding boxcar window that groups the raw spectra into groups of three calculable frequencies. The 90% confidence intervals on the observed spectrum assume six degrees of freedom for each spectral estimate (group of three) shown. The sum of the variance is plotted at the central frequency of the sliding boxcar window. The enhanced power at low frequencies in (a,b) relative to (c,d) is associated with the strong warming trend in both observations and the All-Forcing model runs. There is a strong tendency

for the model spectra to lie within the 90% confidence intervals of the observed spectra, particularly at periods longer than 10 yr (frequency < 0.1 yr^{-1}).

The spectra in Fig 5 (c) and (d) are based on residual time series from observations or model historical runs, where the multi-model ensemble surface temperature time series from the 20C3M volcanically forced historical runs is first subtracted from the observed global mean temperature series or from the individual model historical runs to form residual time series. As a result of this filtering procedure, most of the long-term warming trend (e.g., Fig. 4 a, c) is removed from the time series. The agreement between variance spectra of model and observed residual time series in Fig. 5 (c,d) is not as good as for the original unfiltered spectra (Fig. 5 a,b), particularly for the CMIP3.

Overall, the results of these comparisons suggest that the model simulations have a plausible representation of variability of the climate system, in terms of the spatial pattern of variability and the direct comparison of the time series of observed and historical run global mean surface temperature. The spectral results suggest that the models, particularly the CMIP3, may have some shortcomings in global low-frequency variability simulations, although there are uncertainties in estimates of the internal climate variability as obtained by creating observed residual time series. Overall, these findings encourage us to use the models to assess surface temperature trends at the regional scale in the following sections, with the caveat that there is likely room for improvement in the model simulations of internal variability. Further tests of low-frequency variability are presented in Section 6.

5. Trend assessments: global mean and regional time series

a. Methodology for the "sliding trend" analysis: CMIP5 models

In this section we compare the observed and simulated historical (20C3M) temperature trends obtained from global or regional averages, to assess whether a linear trend signal has emerged from the "background noise" of internal or natural climate variability, as estimated by the models. The primary focus is on the seven CMIP5 models that have Natural-Forcing-Only runs extending to 2010. (In the case of FGOALS-g1.0, the natural-forcing run extends to only 2009, so 2010 is treated as missing.) While we can extend All-Forcing runs to 2010, when necessary, using RCP4.5 projections, this is not tractable for the Natural-Forcing-Only runs. We can use these seven CMIP5 model runs together to assess whether the observed trends have emerged from the background of natural variability and whether they contain an attributable anthropogenic component. We also examine the full sample (23 models) of CMIP5 runs for our All-Forcing run vs. control run analysis. For these 23 models and for the eight CMIP3 models that include volcanic forcing (but for which we generally do not have Natural-Forcing-Only runs), we can ask a more limited set of questions, namely whether the linear trend signal to 2010 in the observations has emerged from the background of internal climate variability and whether the All-Forcing run trends are consistent with the observed trends.

We assess the trends across a wide "sliding range" of start years beginning as early as 1861 and for each of the regions shown in Fig. 6. All trends in the analysis use 2010 as the end year. The general procedure we use is illustrated in Fig. 7 (a) for global mean surface temperature. The

black shaded curve in the figure shows the value of the linear trend in observed global mean temperature for each beginning year from 1880 to 2000, in each case with the trend ending in the year 2010. The HadCRUT4 observed data set contains an ensemble of 100 estimates, and these are used to create an ensemble of observed trend estimates. The black shading depicts the 5th to 95th percentile range of this ensemble. The first year plotted for global mean temperature was 1880 because the areal coverage and temporal coverage requirements for a trend to 2010 were reached in that year. The observed temperature trend to 2010 is about 0.5^{o}C/100 yr (0.05 ^{o}C/decade) beginning early in the record (late 1800s) and increases to about 2^{o}C/100 yr (0.2^{o}C/decade) by around 1980. The observed trend has decreased for more recent start dates, falling below 1^{o}C/100 yr (0.1^{o}C/decade) for trends beginning in the late 1990s.
The blue curve in Fig. 7a shows the "mean of ensemble mean trends" for the Natural-Forcing-Only runs of the seven CMIP5 model subset (see caption). Each of the seven models is weighted equally in the mean of ensemble means, even if a modeling center provided a greater or smaller than average number of within-model ensemble members. The light blue shading in Fig. 7 (a) shows the 5th to 95th percentile range of trend values for the Natural-Forcing-Only runs, which is constructed using the long-term drift-adjusted control run variability (Fig. 1 c,d) from each model. Under an assumption that internal variability in the control run is not substantially different from that in the forced runs, we can use the long control run for each model to estimate the component of inter-realization uncertainty that would be present in the forced trends; this is helpful, since most centers did not provide enough ensemble members to precisely assess this component of the uncertainty.

To prevent any one model from dominating the analysis, our approach also attempts to weight the various models roughly equally. Thus even if one modeling center provided a much longer control run than the others, each of these models would still get an equal weighting in constructing a multi-model sample of internal climate variability. Control runs from each of the seven CMIP5 models contribute equally to the multi-model sample from which the percentile range is constructed, as long a particular model control run is "eligible" for use, meaning here that the length of the usable part of the control run is at least three times the length of the observed trend being examined.

Each randomly selected control run trend (from the seven models used) is combined with that model's ensemble-mean Natural-Forcing-Only trend for that trend length, thus creating a distribution of historical Natural-Forcing-Only trends that includes the uncertainty due to both internal variability and the spread of forced responses across the seven models. The blue region is the 5th to 95th percentile range of this distribution of trends, and thus relates to the uncertainty of single ensemble members (which mimics the real world, itself a "single ensemble member"). Therefore, the distribution of trends used to construct the percentile range includes uncertainty due to both the different natural forcings and responses of the individual models, and the uncertainty due to the internal variability as simulated in the control runs. The random resampling approach is necessary because the available control runs for the various models are of different lengths and yet we purposely chose to give each available model an equal "vote" in estimating internal variability. The samples are drawn from the control runs in the form of 150-yr samples with randomly chosen start dates, with each sample masked with the observed mask of missing data over the period 1861-2010 to create data sets with missing data characteristics that are similar to those of the observations. The analysis in Fig. 7 (a) shows that observed global temperature trends-to-2010 of almost any length are detectable compared to the CMIP5

Natural-Forcing-Only runs and simulated internal variability—even for trends as short as those beginning around 1990. Note that the spread of uncertainty expands for shorter trends, reflecting the fact that the model can internally produce relatively larger-magnitude trend *rates* over relatively short periods.

The dark red curve and light pink shading in Fig. 7 (a) depict the inter-model mean of ensemble means and the 5th to 95th percentile uncertainty range for the All-Forcing runs (i.e., natural and anthropogenic forcings combined) and control runs for the seven-model CMIP5 subset. These are constructed in an analogous way to the Natural-Forcing-Only curves and blue shading, and thus depict the uncertainty due to both internal variability and to the different models' responses to historical climate forcing agents (All Forcings, in this case). The violet shading in the plot is the region where the pink and blue shading overlap, indicating that the 5th to 95th percentile ranges of the All-Forcing and the Natural-Forcing simulated trends at least partially overlap.

In Fig. 7 (a), the black (observed) curve is always within the pink- (or violet-) shaded region, meaning that global mean temperature trends are not significantly different from the CMIP5 historical All-Forcing run ensemble on any time scale, including the most recent 'weak trends' beginning in the late 1990s.

When the black-shaded curve in Fig. 7a lies entirely within (or above) the pink-shaded region and entirely outside of the blue-shaded region, we conclude that the trend from that point to 2010 has a detectable anthropogenic component. Given that the observed global mean surface temperature trends with start dates through about 1990 lie within this region of the graph, we conclude that the observed global surface temperature warming to 2010 is at least partially attributable to anthropogenic forcing according to these model data and observations. Inspection of Fig. 7a further indicates that this detection and attribution result is sufficiently strong that the uncertainty associated with the combined effects of internal climate variability, uncertainty in the model responses to natural forcing, and the uncertainty in the observed ensemble could be a factor of two larger than shown here and the same conclusion would still hold.for start dates from the late 1800s to about the mid-20th century. Our attribution conclusion for anthropogenic forcing and global mean temperature is not as strong as in IPCC AR4 (Hegerl et al. 2007), partly because we are not focusing in this study on quantifying the magnitude or fractional contribution of the anthropogenic forcing. Also, our technique does not use information in spatial patterns to distinguish between different forcings or to quantify the effect of individual forcings (e.g., greenhouse gases). Rather, our focus is on evaluating the evidence for detectable and attributable net anthropogenic influence on surface temperature in various regions around the globe, using the 'best estimates' as provided by current models (without any rescaling). We essentially compare two alternative hypotheses (natural and anthropogenic forcings vs. natural forcings only) and focus down even to the scale of individual 5^o x 5^o grid boxes, which is important for regional climate change assessment.

There are some important caveats to the approach that we use, aside from the obvious one that we rely on models to estimate the internal climate variability levels (which are compared to an derived observed estimate Obs. St. Dev.* in Section 3b). The limited number of ensemble members for the individual models means that there is additional variance in the grand distributions of trends (i.e., pink- and blue-shaded regions) due to our imperfect knowledge of each model's forced response. However, the net impact of this limitation on the spread of the total distribution is a complicated function of several factors. These include the following four

factors: 1) the number of ensemble members a particular model has (which we now show in Fig. 1; the larger the number of ensemble members, the smaller the overestimate of variance); 2) where the models with few ensemble members sit in the distribution (if they are close to the outer edge, the overestimate can be greater than if they are near the middle of the distribution); 3) the variance of the model with few ensemble members or that sits at the outer edge of the distribution; and 4) the relative size of the spread of the individual model ensemble responses vs. the internal variability of the models near the outer edge of the distribution.

We can also estimate an upper limit on the overestimate of the standard deviation, based on the number of ensemble members we use, as about 15-40% at most, with the worst case being for a single ensemble member, where the variance is as much as doubled, so the standard deviation is 40% overestimated. However, given the four factors mentioned above, the effect will typically be considerably smaller than this.

It is also worth noting that the effect of an overestimation of variance in our framework is to make trends too difficult to detect (compared to internal variability or to the internal variability plus natural forcing), but to also make it too easy for All-Forcing trends to be consistent with observations.

We could in principle attempt a simulation to essentially estimate confidence intervals on our confidence intervals, but these would be situation dependent and would vary for different locations around the globe, time period, etc. We have chosen to leave this extension for further studies, but note that the above issues should be considered in evaluating our results.

b. Detection/attribution findings for various regional indices

The sliding trend/ detection and attribution analysis discussed above for global mean temperature can be applied to various regions around the globe (Fig. 6). Here we briefly summarize the findings of such an application (panels shown in Figs. 7 and 8).

1) MAJOR LARGE-SCALE REGIONAL INDICES

For **global sea surface temperature (SST)** (Fig. 7b), trends to 2010 are clearly detectable for starting years up to about 1990. The observed trends are only marginally attributable to anthropogenic forcing for trends beginning around the mid-20th century, otherwise an attributable anthropogenic signal is clearly apparent for the detectable trends. For **global land surface temperature** (Fig. 7c), an attributable anthropogenic signal is clearly seen in the observed trends for all start dates from about 1885 up to about 1990, so the case for attribution is similar to that for global sea surface temperature. The anthropogenic warming signal is so much stronger over land than over ocean, that it readily detectable and attributable despite the greater intrinsic variability over land than over ocean. **Northern hemisphere temperature** (Fig. 7d) roughly mirrors the results for the global temperature indices, with robust detection and attribution for start years up to about 1990. **Southern hemisphere temperature** (Fig. 7e) results are similar though not quite as robust as for the Northern hemisphere, as the start dates with attributable anthropogenic influence extending up to about 1980, rather than 1990.

The **northern hemisphere extratropics** (30^{o}-90^{o}N) series (Fig. 7f) has robust detection and attribution up to around a 1990 start date. However, warming of the **southern hemisphere extratropics** (30^{o}-90^{o}S; Fig. 7g) is slightly less robust than the northern hemisphere, as detection/attribution extends to starts dates up to about 1980. The trends for the **southern extratropics** are relatively constant over a range of start dates from 1900 to 1970, in contrast to **northern hemisphere** series which shows a period of higher warming trend rates for trends to 2010 beginning in the second half of the 20^{th} century. The **southern extratopics** trends from 1900 are marginally consistent with the All-Forcing model trends, as they are near the upper edge (95^{th} percentile) of the modeled distribution. An interesting feature of many of the global and regional trends in Fig. 7 is that there is essentially no start date for which the 5^{th} to 95^{th} percentile range of the All-Forcing and Natural-Forcing-Only simulated trends are not at least partially overlapping. That is, in some sense the All-Forcing and Natural-Forcing trends from the models are typically not completely distinguishable from each other. The same will be true for many of the subsequent regional series analyzed, and particularly for land regions and ocean regions with pronounced multi-decadal variability. **Tropical surface temperatures,** which combine land and ocean (Fig. h) regions, show robust detection and attribution for trends to 2010 with start dates as late as the 1970s.

2) REGIONAL SEA SURFACE TEMPERATURE INDICES

Tropical SST's (20^{o}N-20^{o}S; Fig. 7i) show similar robust detection and attribution results (for start dates as late as about the 1970s) to those for the tropical surface temperature as a whole. **Indian Ocean SST**s (Fig. 7j; see Fig. 6 to identify region IO) exhibit robust detection and attribution for start dates up to about 1990, despite a larger observational uncertainty, particularly for trends beginning from the 1940s through the 1980s. A similar result is seen for the **tropical Indian Ocean/western Pacific** warm pool index (Fig. 7k) and for the **tropical west Pacific** (Fig. 7l), which are important regions as they are dominant large-scale regions for tropical convection; these have a detectable anthropogenic component for trends beginning up until about 1990 and 1980, respectively. The **tropical east Pacific** (Fig. 7m) shows a detectable anthropogenic component for trends to 2010 beginning from about 1900 to about 1920. However, trends beginning from 1920 to 1970 are only marginally detectable as the black region (observations, including uncertainties) is not clearly outside of the blue/violet (natural forcing) region. **North Pacific SSTs** (25^{o}-45^{o}N, Fig. 7n, see Fig. 6 to identify region), have a detectable anthropogenic component but only for start dates up to about 1910. A marginally detectable signal is found for a narrow range of start years in the 1970s. Otherwise, the trends are not detectable according to our analysis.

We analyzed four separate regions of the Atlantic Ocean, a basin is noted for pronounced multi-decadal variability. In the **South Atlantic** (Fig. 7o), there is a detectable anthropogenic warming for start dates up to the late 1970s. An interesting feature in this region is that warming trends from the late 1800s are near the 95^{th} percentile of the all-forcing model simulations. **North Atlantic SSTs** (45^{o}-60^{o}N; Fig. 8a) exhibit no detectable trends outside of the range of natural variability for any start dates, according to our analysis. This region is notable for having probably the least detectable signal of any of our study regions around the globe. Despite the lack of detectable trends, the observed trends are at least consistent with the All-Forcing runs, which have a very wide 5^{th} to 95^{th} percentile range of trends due to the large simulated internal variability, as will be shown later in this section. In the **subtropical north Atlantic** (20^{o}-45^{o}N;

Fig. 8b) an anthropogenic signal is detected for start dates from about 1890 to 1920 and around 1970, but otherwise is only borderline detectable up to about 1980. In the **tropical North Atlantic "main development region"** for Atlantic tropical cyclones (Fig. 8c), there is a detectable anthropogenic warming to 2010 for start dates up to about 1960, and then only intermittently for start dates up to about 1990.

MAJOR LAND REGION TEMPERATURE INDICES

We now summarize the characteristics of surface temperature trends in major continental regions, beginning with Eurasia, Africa, and Australia. The **Europe** temperature index (Fig. 8d) has detectable anthropogenic warming trends for start dates up to about 1990, as the observed trends (even accounting for observational uncertainty in the HadCRUT4 data set) are outside of the range of the Natural-Forcing trends but lie well within the range for the All-Forcing trends. The **Africa** index (Fig. 8e) has detectable anthropogenic warming trends for start dates up to the last start year analyzed (2001). Our analysis of **African** temperature trends only extends back to start dates beginning in the mid-1920s, due to more limited data coverage. For **northern Asia** (Fig. 8f), our start dates extend back to the early 1900s and show a clear detectable anthropogenic warming signal for start dates extending from there up to about 1980. For **southern Asia** (Fig. 8g) our analysis shows a similarly strong detectable anthropogenic warming signal for start dates extending from the late 1800s through about 1990. An interesting feature of the **African** and **southern Asia** results is that the 5^{th} to 95^{th} percentile range of the All-Forcing trends (pink) from much of the 20^{th} century is much wider than the range for the Natural-Forcing runs. Since the contribution from internal variability (estimated from the control runs) is the same for the All-Forcing and Natural-Forcing trend results, the uncertainty range of the All-Forcing ensemble mean trends across the models must be comparable to or substantially larger than the uncertainty due to internal climate variability alone. The **Australia** temperature index (Fig. 8h) shows detectable anthropogenic warming trends for start dates from the late 1800s to about 1970.

Considering now the land regions of North and South America, the index for **Canada** (Fig. 8i) shows detectable anthropogenic warming trends for start dates up to the late 1970s. In contrast, for the **Alaska** index (Fig. 8j), a detectable anthropogenic warming trend to 2010 is most clear for start dates over the more limited range of about 1940-65. Trends for post-1970 start dates are generally not detectable. For the **continental United States** (Fig. 8k) an anthropogenic warming trend to 2010 is detectable for start dates of about 1900 to 1975. For start dates of about 1860 to 1900, the warming signal is not clearly detectable. The temperature index for **Mexico** (Fig. 8 l) indicates that observational uncertainties play an important role for detection and attribution results in this region. A detectable anthropogenic warming trend is seen for start dates of about 1910-1920 and about 1965-1980, otherwise the trends are not detectable. In contrast, for the **South America** index (Fig. 8 m), the temperature trends to 2010 are mostly detectable for start dates from about 1910 to 1950 and around 1970, but these trends are not necessarily attributable to anthropogenic forcing for these periods because the observed trend range is not entirely within the pink region (range of All-Forcing simulated trends). Rather, they appear systematically smaller than the simulated trends, after accounting for observational uncertainties. Anthropogenic warming trends to 2010 are not clearly detectable for the **South America** index for any start years examined.

Temperature trends for the **southeastern United States** index (Fig. 8o) are of particular interest because the trend behavior in this region is different from most other land regions around the globe, as has been pointed out in a number of previous studies (e.g., Knutson et al. 1999, 2006; Portmann et al. 2009). According to our present analysis, trends to 2010 in this index are detectable only for a limited range of start years (mid-1950s to the mid-1970s). For that limited set of start years, an anthropogenic warming trend to 2010 is detectable in our analysis. The trends in the index to 2010 at least are consistent with All-Forcing runs for all start years after about 1940, but the warming trends even after 1940 are for the most part not strong enough to be detectable against the background of natural forcing and internal climate variability. This behavior contrasts with the index for the **rest of the continental United States** (that lies outside of the southeastern U.S.) (Fig. 8 n), where an anthropogenic warming trend to 2010 is broadly detectable for start years ranging from about 1890 to the mid-1970s.

c. Consistency test findings using CMIP3 and CMIP5 models

Our regional temperature indices analysis in subsections 5(a) and 5(b) (i.e., Figs. 7 and 8) focused on the subset of seven CMIP5 models that had Natural-Forcing-Only runs that extended to 2010. Here we conduct a complimentary assessment (for a more limited set of regions) that compares these results with similar analyses for the eight CMIP3 models (All-Forcing and control runs) and with the full set of 23 CMIP5 models (All-Forcing and control runs). Where necessary, the All-Forcing 20C3M runs were extended to 2010 using A1B (CMIP3) or RCP4.5 (CMIP5) projection runs; this procedure was not tenable for the Natural-Forcing-Only runs due to the strong differences in forcing between Natural-Only and the A1B or RCP4.5 scenarios for the extension years to 2010. Our analyses for the CMIP3 models (and the 23 CMIP5 models as shown in the middle column of Fig. 9) therefore only compare internal climate variability (control runs) with All-Forcing historical runs. Thus, we cannot use these results to draw firm conclusions about detection of anthropogenic trends, because the alternative hypothesis (Natural-Forcing) is not available through 2010 for all of the models. Nonetheless, we can draw some conclusions about detection of significant trends (against a background of internal climate variability) and about consistency of observed trends versus the trends in the All-Forcing 20C3M experiments.

Our procedure is illustrated for the **global temperature** analysis in the top row of Fig. 9 (a-c). Figure 9c is identical to Fig. 7a and is repeated here for reference only. Figure 9a shows the 5^{th} to 95^{th} percentile range for the observed trends to 2010 (black shading); the 5^{th} to 95^{th} percentile range for the All-Forcing runs from the eight CMIP3 models (pink shading, with the red curve depicting the ensemble mean); and the 5^{th} to 95^{th} percentile range of control run trends from the same eight CMIP3 models (green shading). Violet shading illustrates regions of overlap of the pink- and green-shaded regions. Where the black curve lies outside of the green- or violet-shaded region, the observed trend is detectable compared to internal climate variability in the CMIP3 runs. Where the observed curve lies within the pink (or violet) shading, the observed trend is assessed as consistent with the CMIP3 All-Forcing ensemble of runs.

Figure 9a (CMIP3) indicates that the observed **global mean temperature** trends to 2010 are detectable (inconsistent with internal climate variability in the eight CMIP3 models) for start dates from about 1880 to the mid-1990s, and are consistent with the CMIP3 All-Forcing run trends to 2010 for essentially all start dates from 1880 to 2000. Similar conclusions are evident for the 23 CMIP5 models as shown in Fig. 9b. As noted earlier, similar results are seen for the

seven CMIP5 models when we incorporate the Natural-Forcing-Only runs in the tests (Fig. 9c), although there the detectability of the observed trend extends to start dates as late as about 1980, rather than into the mid-1990s.

For **tropical SST** (Fig. 9d-f) the CMIP5 models, including the seven model subset with Natural-Forcing-Only runs to 2010 (Fig. 9 f), indicate robust detection and attribution for trends to 2010 for almost all start dates as late as about the late 1970s, as discussed earlier. The consistency with the All-Forcing runs (all 23 CMIP5 models) is only marginal for a period of start dates around 1960. A similar consistency result is seen for the 23 CMIP5 models (Fig. 9e) where we compare their All-Forcing runs with their control variability. The observed trends to 2010 appear to be detectable against the internal variability (control run) background of the 23 CMIP5 models for start dates as late as about 1990. For the eight CMIP3 models (Fig. 9d), the observed trends to 2010 are detectable for start dates up to 1990, similar to the CMIP5 models (Fig. 9e). However, the eight CMIP3 All-Forcing runs are not as consistent with the observed trends to 2010 as the 23 CMIP5 All-Forcing runs. In fact the CMIP3 All-Forcing runs appear only marginally consistent with the observed trends to 2010 for most of the start dates from 1880 through about 1980. This illustrates that the relatively modest levels of estimated internal variability in this basin lead to a strongly detectable warming signal, but also make it difficult for a model to be assessed as consistent with the observations, as the margin for error is relatively small.

The **North Atlantic** (45^{o}-65^{o}N) was highlighted earlier as a region with no detectable trends compared with the CMIP5 Natural-Forcing-Only runs and internal climate variability combined (Fig. 9i). This is perhaps not surprising, given the substantial intrinsically-generated fluctuations on multi-decadal time scales in this region (see e.g. Yang et al. 2013). We see from the green and violet shaded regions in Figs. 9 g,h that the range of trends to 2010 due to internal climate variability alone in the CMIP3 and CMIP5 models is quite large and appears to largely account for a similar wide range of simulated trends in the All-Forcing runs. This also helps allow the observed trends to 2010 to be consistent with the CMIP3 and CMIP5 All-Forcing trends for all of the start dates examined, despite the fact that the observed trends are not detectable (i.e., not distinguishable from control run variability alone).

For the **southeastern United States** index (Fig. 9 j-l) there is slightly more evidence for detectable trends to 2010 versus the internal variability samples in Fig. 9 j,k (start years 1950 to 1980) than versus the combined Natural-Forcing/internal variability sample of trends from the seven CMIP5 models (blue shading in Fig. 9 (l)) with the latter having only marginally detectable trends and only for start dates from the mid-1950s to the mid-1970s). For start years prior to about 1940, the distribution of observed trends lies near the edge and even outside of this 5th to 95th percentile range for the All-Forcing runs (pink/violet shaded envelopes), especially for the CMIP3 model sample (Fig. 9j). We thus conclude that even accounting for internal variability, the CMIP3 and CMIP5 historical runs trends-to-2010 tend to be inconsistent or only marginally consistent with the observed southeastern U.S. surface temperature trends, particularly for starting dates in the early 20^{th} century. This means that the CMIP3 and CMIP5 All-Forcing runs can be falsified, at least for this relatively small region, and further implies that there remain as yet unexplained discrepancies between the historical simulations and observations for trends in this region. We note that our tests are conducted on a large sample of at least partly independent regions, and thus we would expect some fraction of the area to have

values that are too high or low due to chance. Further discussion of this issue in the context of "global significance testing", can be found, for example, in Knutson et al. (1999).

The results for the **rest of the continental United States** index (outside of the southeastern United States; Fig. 9 m-o) are fairly consistent between the CMIP3 (m) and the CMIP5 models (n, o), although as discussed above, the nature of our conclusions are different for Fig. 9 (m and n) than for Fig. 9 (o), with the latter one including also the ensemble mean and additional uncertainty range associated with the different model responses to Natural Forcings.

6. Grid point-scale detection and attribution tests

a. Multi-model ensemble assessment

1) 1901-2010 TRENDS

The procedures in Section 5 that were used to categorize observed trends at individual grid points as detectable, attributable in part to anthropogenic forcing, consistent with All-Forcing runs, etc. can be applied at the grid-point scale, and the categories displayed in map form, for a selected trend period. For example, Fig. 10 shows the results of such a category analysis for the observed vs modeled trends for 1901-2010, with the bottom row showing category maps for the CMIP3 All-Forcing runs (e) and CMIP5 All-Forcing and Natural-Forcing-Only runs (f). The linear trend maps for observed temperature (1901-2010) and the CMIP3 and CMIP5 All-Forcing ensemble means are shown in Fig. 10 (a-d) for reference. The observed trend map shows broad-scale warming trends since 1901 at almost all locations around the globe, with areas of cooling in only a few regions, mainly in the high latitude North Atlantic and the southeastern United States. The CMIP3 and CMIP5 multi-model ensemble trends show broadly similar magnitude and pattern of cooling to observations, where the agreement can be quantitatively tested by our consistency tests as described in the previous section. For the tests described in this section, we use only the ensemble mean observed trend and thus do not consider observational uncertainty, which was examined in the previous section.

Figure 10 (f), for the seven CMIP5 models with both All-Forcing runs and Natural-Forcing-Only runs to 2010, builds upon the regional time series analysis shown in Figs. 7-8. The white regions in Fig. 10 (f) indicate where the observed trend is not detectable compared to the Natural-Forcing-Only runs (where the uncertainty estimates incorporate both simulated internal climate variability from the seven control runs and uncertainties in the Natural-Forcing-Only ensemble mean). The dark grey regions in Fig. 10 (f) do not have sufficient data coverage for our tests. To determine if a grid point has "sufficient coverage" to include in our maps and analyzed area, we divide a given trend period (e.g., 1901-2010) into five roughly equal periods, and require that each of the five periods has at least 20% temporal coverage for annual means, where an annual mean is considered available if at least 40% of the months are available for the year. The various colored (non-white, non-grey) regions in Fig. 10 (f) indicate where the trends are detectable, with the category identified on the legend. The yellow-orange regions show where the warming trend is detectable but still less than the lower end (5^{th} percentile) of the All-Forcing trend distribution. The light-red and dark-red regions indicate where the observed trend has a detectable anthropogenic component; for the darkest red regions the observed warming trend is so large that

it exceeds the 95th percentile of the modeled distribution, but here we still interpret this as implying a detectable anthropogenic component. For cooling trends (blue regions), we have analogous terms to those used for the various warming cases, although these cases are almost absent for the 1901-2010 trends in our analysis.

The results for Fig. 10 (f) show that most of the global area with sufficient temporal coverage is categorized as having attributable anthropogenic warming (either consistent in magnitude or significantly larger than in the CMIP5 All-Forcing runs). The larger-than-simulated warming trends occur in a few regions in the extratropical South Pacific, the South Atlantic, the far eastern Atlantic and the far western Pacific. In only a relatively small percentage of the globe is the observed trend classified as not a detectable change (white regions in Fig. 10 f). These include mainly the mid- to high-latitude North Atlantic, eastern United States, parts of the eastern tropical and subtropical Pacific and the North Pacific.

A similar analysis for the CMIP3 All-Forcing runs (eight models with volcanic forcing) is shown in the left column of Fig. 10 (a,c,e). The category names for the assessment (Fig. 10 e) are different than for the CMIP5 models (Fig. 10 f) because a Natural-Forcing-Only ensemble is not available in the archive for the CMIP3 models. Therefore, our categories for CMIP3 (see legend) are limited to assessing consistency, either with the internal variability of the control runs or with the All-Forcing runs, and we do not address the question of attribution to anthropogenic forcing. The observed widespread warming trends shown in Fig. 10 (a) are assessed as detectable (compared with CMIP3 control run or internal climate variability) over most of the global region with sufficient coverage. Only in some regions of the North Atlantic, eastern United States, and North Pacific (white regions in Fig. 10 (e) is the observed trend not detectable. In only a very minor fraction of the analyzed area is there a detectable cooling trend since 1901 (blue shading in Fig. 10 e), according to our analysis. Yellow-orange regions (where the warming trend is detectable but less than simulated) occur in some regions of the lower latitudes, and are more common in the CMIP3 assessment than the CMIP5 assessment. Regions with significantly greater than observed warming trends (dark red) tend to occur in regions outside of the deep tropics for the CMIP3 assessment (Fig. 10e), as was also found for the CMIP5 assessment (Fig. 10f).

2) 1951-2010 TRENDS

Figure 11 explores how the results seen for 1901-2010 trends in Fig. 10 are altered when we analyze the trends for 1951-2010. The observed trend map (Fig. 11 a) shows a more spatially varying structure than the trend map for 1901-2010 (Fig. 10 a). The Asian and North American extratropical land regions have warmed substantially more than oceanic regions since 1951. This amplification of warming over land since 1951 is also evident in the All-Forcing 20C3M ensemble means for both the CMIP3 eight-model set (Fig. 11c) and the CMIP5 seven-model (Fig. 11d)—although the contrast between the continental and oceanic regions is more pronounced in the observed trend map than in the multi-model ensembles. The category maps (Fig. 11 e, f) show a few regions with dark-red shading (observed warming significantly greater than simulated).

The observed trend map (Fig. 11 a, b) shows a region of notable cooling over the mid-latitude North Pacific and a smaller region of cooling trends in the high-latitude North Atlantic just south of Greenland. These cooling regions are assessed as having no detectable change (Fig. 11e, f),

meaning that the cooling trends lie within the 5th to 95th percentile range of the simulated trends from the model control runs (CMIP3) or combined control run/Natural-Forcing runs (CMIP5). Non-detection of trends for 1951-2010 (white category, Fig. 11 e,f) are also found over large regions of the North Pacific, the central equatorial Pacific, the mid- to high-latitude North Atlantic, the far Southern Ocean near Antarctica, and in a few scattered continental regions such as the south-central or southeastern United States.

Figure 11 (f) indicates where observed trends (1951-2010) are attributable, at least in part, to anthropogenic forcing (light-red and dark-red regions). These regions cover most of the global area that has detectable trends, and for the 1951-2010 trends are comprised predominantly of regions where the trends are consistent with the All-Forcing ensembles (i.e., light red). Minor parts of eastern Asia have warming trends that are significantly larger than simulated in the CMIP5 All-Forcing runs (dark-red shading). The category results for the eight CMIP3 models (Fig. 11 e) are generally similar overall to those for the CMIP5, although the categories in Fig. 11 (e) do not include attribution to anthropogenic forcing (see legend), since the CMIP3 set of models does not include Natural-Forcing-Only runs that are necessary for such an attribution.

Regions in Fig.11 (e, f) with warming trends that are detectable but significantly less than simulated in the All-Forcing runs (yellow-orange regions) are quite rare in our analysis.

3) 1981-2010 TRENDS

The trend assessment results for the much shorter period 1981-2010 are presented in Fig. 12. The observed trend map (Fig. 12 a) has much more spatial structure than for either of the longer trend periods in Figs. 10a and 11a. Since 1981 there have been extensive regions of cooling trends over the tropical and subtropical eastern Pacific, Gulf of Alaska, and much of the high latitude Southern Ocean. The trend assessment (Fig. 12 e, f) shows that for the most part, the cooling trends in these regions are not detectable. In fact, since less than 5% of the globe has "detectable" cooling trends, the percent of occurrence of the blue regions is not significantly different from what could occur from sampling variability alone.

The large expanses of the globe without detectable trends (1981-2010) in Fig. 12 contrasts with the earlier finding of detectable warming in most analyzed regions for the longer trend analyses (Figs. 10, 11). The loss of a detectable signal, as one proceeds to later start dates in the 20th century--and shorter trend periods--is not unexpected. For example, the results in Figs. 7-9 showed how the trend *rates* for internally generated trends in the model become higher for shorter trend periods, as the models can produce strong internally generated trend rates over relatively short periods. Comparing the category maps for different start dates (Fig. 10-12), the loss of detectability, as one proceeds to later start dates, occurs first in the extratropical North Atlantic (north of 40°N) and over large parts of the North Pacific, extending into the tropics, as seen for the 1951-2010 trends (Figs. 11). For the late 20th century start dates (e.g., 1981-2010; Fig. 12) the region of no detectable warming expands to cover most of the southern oceans, south of 40°S, and extending south from 20°S in the South Atlantic. This non-detection region also expands to include most of the eastern tropical and subtropical Pacific and much of the northern extratropics over Eurasia, North America, and the North Pacific.

Of the regions with detectable trends for 1981-2010 (Fig. 12 e, f), the vast majority of grid points have trends that are consistent with the models (light red) and thus are at least partly attributable

to anthropogenic forcing (CMIP5; Fig. 12f) or, in the case of the CMIP3 models (Fig. 12 e), at least consistent with All-Forcing runs. These areas include large regions of the tropics, subtropics, and mid-latitudes within about 40-50 degrees of the equator particularly in the Indian Ocean, western Pacific, South Asia, and the tropical Atlantic. The relatively robust emergence of a significant warming signal over a relatively short time period (30 years) in the lower latitudes, as in Fig. 12 (f), is reminiscent of the recent study of Mahlstein et al. (2011), who conclude that the earliest emergence of significant greenhouse warming will occur in the summer season in low-latitude countries. They examined land regions and looked at signal emergence for particular seasons (whereas we examine land and ocean regions and focus on annual means). However, both studies point toward early emergence of anthropogenic warming signals in lower latitudes, as opposed to most high latitude continental regions. Some exceptions we note in Fig. 12 (f) include the significant anthropogenic warming trends (1981-2010) in the vicinity of Greenland and in some land regions near the edge of the Arctic Ocean.

The rare occurrence of yellow-orange area (which in our convention designates warming that is detectable but significantly less than simulated) on the assessment maps for 1981-2010 (Fig. 12 e, f) can be explained by referring to the sliding trend analyses in Figs. 7-9. The unshaded area on those graphs between the pink- and blue-shaded "envelopes" corresponds to detectable warming that is less than simulated. However, this region typically systematically shrinks as one progresses to later start dates. That is, for shorter trend periods, it becomes much more difficult to distinguish the simulated All-Forcing trend distribution from the trend distribution of the Natural-Forcing-Only runs (CMIP5) or from the control runs (CMIP3).

4) ENSEMBLE MEAN ASSESSMENT STATISTICS ACROSS TIME

In Fig. 13, we explore how the percent of analyzed area with various category classifications changes for different start years (all for trends ending in 2010). Figure 13(b) shows the aggregate percent area results for the CMIP5 models, using the seven models that have Natural-Forcing-Only runs extending to 2010. The total percent of analyzed area (i.e., regions with sufficient data coverage) that was assessed as having attributable anthropogenic warming trends (black curve) was about 80% for trends over the period 1901-2010. This drops from about 70% to 55% for start dates increasing from 1931 to 1971, before dropping sharply to about 25% for the shortest period (1991-2010). The rapid decrease in percent of area with attributable anthropogenic warming beyond the 1971 start date is apparently due to the temporary pause in global warming from about 1940 to 1970, which was preceded by a relatively strong rate of global warming during early 20th century (Delworth and Knutson 2000). The end of this pause, around 1970, is a time period during which the prospects for detection of a warming signal are at least temporarily enhanced against a backdrop of a gradually declining percentage as the start date is moved forward through the 20th century. The blue curve in Fig. 13b (percent of analyzed area with no detectable change) shows generally opposite behavior to the black curve, increasing from a low of about 20%, for 1901-2010 trends, to a high point of over 75% for the latest start period analyzed (1991-2010). The analysis thus illustrates the advantages of a long record for detectability of the warming trend. The red curve shows that roughly 15 % of the analyzed area has warming that is detected but greater than simulated, for start dates around 1900. This percent of area with trends that are attributable to anthropogenic forcing but significantly greater than simulated also diminishes as the start dates move later in the century, possibly because of the

growing width of the simulated trend distributions associated with internal climate variability, implying that it becomes difficult for an observed trend to be large enough to be inconsistent with the All-Forcing distributions on the high side.

Figure 13 (a) summarizes the comparison between the CMIP3 (eight-models with volcanic forcing) and CMIP5 (23-model) results (solid lines vs. dashed lines) for various common categories. This figure shows the percent areas corresponding to the maps in Figs. 10-12 (a, c, e) for the CMIP3 models, but for a range of start dates. For the CMIP5, we use results for all 23 models that have volcanic forcing, since a Natural-Forcing-Only experiment (extending to 2010) is not required for the comparisons shown in Fig. 13 (a), and thus we are not limited to the seven-model subset of CMIP5. The percent area where the warming for the period 1901-2010 is detected and either consistent or greater than simulated (black curves) is about 80% for CMIP3 and almost 90% for CMIP5. This percentage decreases for start dates of 1931 or 1941, before rising to a temporary peak of about 70% for the 1961 start date and then falling again for later start dates. As discussed earlier, the temporary rise for mid-century start dates is likely due to the enhanced detectability of trends that start within the "relative trough" or temporary interruption of global warming that occurred around this time following the relative peak in global temperatures around 1940. For start dates up to about 1931, the black curve for the CMIP5 models (dashed) is more than 5% higher on average than the (solid) one for the CMIP3 models. Thus, the 23 CMIP5 model All-Forcing runs appear at least slightly more consistent with observed trends than the eight CMIP3 All-Forcing runs, at least for the case of trends to 2010 starting earlier than 1940. However, for trends with start dates from 1951 through 1991, the CMIP3 All-Forcing runs are about equally consistent with observations. Other features in Fig. 13 (a) are generally similar to those described for the seven CMIP5 models (Fig. 13 b), although the category descriptions (conclusions about attribution) are necessarily different. The general temporal behavior of the various curves through time is quite similar between the solid (CMIP3) and dashed (CMIP5) models in Fig. 13 (a).

b. Model by model trend assessment

In contrast to the analyses in the previous subsection (Figs. 10-13) which focused on the multi-model ensemble means vs. observations, in this subsection we consider the individual models within the CMIP3 and CMIP5 ensembles and assess what percentage of individual models meet certain criteria. That is, the determination of whether a given CMIP3 or CMIP5 individual model is included in a category (e.g., "warming- detectable and consistent") for a given grid point is based on the evaluation of the historical runs and control runs for that model alone. In this section, we also introduce and apply a variance consistency test as an additional consistency test for the models vs. observations.

We will introduce and describe the various tests as we discuss the different panels in Fig. 14, which contains the analysis of the eight CMIP3 models (with volcanic forcing) vs. observations for linear trends over the period 1901-2010. Figure 14 (a) and (b) present the observed and multi-model ensemble mean trend maps for reference; these were discussed earlier for Fig. 10. Figure 14 (c) shows the fraction (or percent) of models, at each grid point, that have no detectable trend. The area-weighted global average of this fraction is 0.09, and the most prominent regions with no detectable trend are in the North Atlantic (south of Greenland), the mid-latitude North Pacific, and the southeastern United States. Figure 14 (d) shows the fraction of models at each grid point with warming that is detectable but less than simulated in the All-

Forcing runs. The global average fraction is 0.22, and the most prominent regions of occurrence are in the tropics, meaning that the eight CMIP3 models, viewed independently, have a tendency to simulate too rapid a century-scale warming in the tropics. The warming is detectable and consistent with the All-Forcing runs for a global average fraction of 0.34 of the models (Fig. 14 e), with a spatial pattern that is fairly evenly distributed around the analyzed areas of the globe. The warming is detectable and significantly greater than simulated for a global average fraction of 0.32 of the models (Fig. 14 f), with the most prominent occurrence of this category being in the mid- to high latitudes of both hemispheres. Warming is detectable for about 89% of the models, on average around the globe (Fig. 14 g)—essentially the inverse of the results in Fig. 14 (c). Warming is detectable and consistent or greater than simulated for two thirds of the models, on average, (Fig. 14 h) which shows essentially the inverse of the pattern in Fig. 14 (d) over most of the globe, and indicates that the simulated warming tends to be too weak in mid- to higher latitudes in the CMIP3 All-Forcing runs. The observed and CMIP3 simulated (All-Forcing) trends are assessed as consistent for 39% of the models on average (Fig. 14 i); this category includes cases where the trend is not detectable, but still consistent with the All-Forcing runs (see Figs. 7-9 for example). This fraction field (Fig. 14i) has a fairly even spatial distribution over the global analyzed area.

One limitation of our approach is that models with unrealistically large internal variability have some advantage over models with more realistic variability, in that it is easier for high-variability models to have trends that are consistent with observations, since the margin of error is greater. To address this concern, here we apply a second test (a variance consistency test) to the models. Then a model that has both a consistent trend and consistent variability, compared with observed estimates, will be ranked more highly in a metric test compared with a model with consistent trends but inconsistent variability. In other words, this expands our consistency tests into a two-dimensional space (trend and internal variability).

The variance consistency test for the eight CMIP3 models with volcanic forcing (Fig. 14 j) is constructed as follows. For each grid point, we estimate the adjusted observed standard deviation of low-frequency (>10 yr) internal variability (Obs. St. Dev.*) as discussed in Section 3b. This variability is compared with a distribution of low-frequency standard deviations from the model control run, which is obtained by drawing 50 random 110-yr samples of combined SST and surface air temperature from the drift-adjusted control run (see Section 3 a), masking missing data periods with the observed mask for the given grid point, low-pass filtering, and computing the 50 standard deviation estimates. If the Obs. St. Dev.* value for the grid point lies within the 5^{th} to 95^{th} percentile range of the combined control run distribution, the model is assessed as having low-frequency internal climate variability that is consistent with the observations according to this test. There are important limitations of this test, which we recognize at the outset. When applied to a single model, as done here, a single model's control run may not be long enough to provide an adequate sample of the 5^{th} to 95^{th} percentile range of low-frequency (>10 yr) variance estimates; indeed, this is an important reason to advocate for longer control runs (or larger ensemble sizes) in future CMIP designs. In addition, the observed residual, which is needed for comparisons with control run variability, has some uncertainties, as the multi-model ensemble mean forced response only approximately removes the forced climate signal from the observations. Our adjustment procedure used to create Obs. St. Dev.*, described in Section 3b, attempts to account for this uncertainty.

Figure 14 (j) illustrates the results of applying the test. On average, 26% of CMIP3 models have variability consistent with observations, according to the test. Locations where the modeled low-frequency variability is consistent with observations are fairly evenly distributed around the globe, although the fraction is notably low in the southeastern Pacific and south Atlantic basins.

Figure 14 (k) shows the map of the fraction of the CMIP3 models where both the variability and trend are consistent with observations on a grid point basis according to our tests. The global average fraction is 0.11, indicating that achieving consistency with both tests simultaneously at the grid point scale is a challenge for the CMIP3 models.

The variance consistency test can also be applied to the global mean temperature series (Fig. 4b). We find that seven of the eight CMIP3 models (88%) have low-frequency variance for their global mean temperature that is consistent with the observed residual, according to our test, with one model having variance that is significantly too low.

Figures 15 and 16 present the same analysis as Fig. 14, but for the 23 CMIP5 models with All-Forcing runs (Fig. 15), and for the subset of seven CMIP5 models that have at least one Natural-Forcing-Only run extending to 2010 (Fig. 16). The mapped results for the 23 CMIP5 models (Fig. 15) are rather similar overall and have similar spatial features to those for the eight CMIP3 models (Fig. 14) discussed above. One notable difference is that the CMIP5 models in both Fig. 15 and Fig. 16 have a greater global mean fraction of models with consistent low-frequency variance (0.31) than the CMIP3 models in Fig. 14 (0.26). The globally averaged fraction of models that have both consistent trend and variance (panel k) is about the same in CMIP5 (0.12 – 0.13) as in the CMIP3 sample (0.11). Figure 16, for the seven model subset of CMIP5 models, shows where trends are assessed as containing attributable anthropogenic trend contributions. The analysis indicates that the globally averaged percent of the seven CMIP5 models with attributable anthropogenic warming at the grid point scale over the 1901-2010 period is over 70% (Fig. 16 h). The globally averaged percentage of models with both attributable anthropogenic warming and consistent low-frequency variance is 22%, according to the tests described above (Fig. 16 l).

The variance consistency test can also be applied to the global mean temperature series for both the full set of 23 CMIP5 models and the seven model subset of CMIP5 models. This test indicates that 16 of the 23 CMIP5 models 70%), and five of the seven–model CMIP5 model subset (71%), have global mean low-frequency variance that is consistent with observations. This is a smaller fraction than for the CMIP3 models (88%). In cases of inconsistency, the model variance is too low more often than too high (6 low vs. 1 high for the CMIP5 23 models; and 2 low vs. none high for the CMIP5 seven-model subset). For cases other than the global mean, Figures 2 and 3 depict where the low-frequency variability of individual models, or the ensemble-average low-frequency variability across the models, tends to be either too low or too high, compared to the adjusted observed internal variability estimate (Obs. St. Dev.*).

As has been discussed mentioned earlier, there are a number of limitations in our trend variance estimates and consistency tests. We hope to improve on the variance consistency tests in a future study; for example, there are other model-observation comparison paradigms that can be explored (e.g., Annan and Hargreaves 2010). Meanwhile, we stress the need for longer control runs and/or greater numbers of independent ensemble members from the models in order to more robustly assess the various models' low-frequency variability.

Figure 17 summarizes several globally averaged trend consistency metrics as a function of trend start year for the individual models in the CMIP3 and CMIP5 samples. Fig. 17 (b, d, and f) also assess the consistency of the models' low frequency variability, as these include both a trend consistency test and a variability consistency test. In the various panels of Figure 17, we compare, across the models, the fraction of analyzed area where there is both a detectable change in observations and where this detectable change is consistent with the individual climate models. Note that these metrics do not include the fraction of area where a climate model is consistent with observations but there is not a detectable trend.

While all metrics have shortcomings, the particular metrics in Fig. 17 have at least some useful compensation properties. For example, for a model with unrealistically large internal variability, the enhanced potential for consistency of modeled and observed trends due simply to the larger internal variability is partly compensated by a reduction in the area assessed as having detectable trends according to that model. The two-dimensional (trend and low-frequency variance) consistency tests provide for an even greater compensating balance against the potential metric problem mentioned above.

The results in Fig. 17 (a, c) show that the individual CMIP3 and CMIP5 models have rather similar behavior in terms of fraction of globally analyzed area with consistent detectable trends (typically ranging from 20 to 50%). There is somewhat more spread among the CMIP5 models, although there are more models in the CMIP5 sample as well. This trend consistency metric tends to reach a peak value around 1960-1970 start dates before declining for later start dates, for reasons discussed for Fig. 13. When a variance consistency test is added (Fig. 17 b,d), the percent of analyzed global area with both consistent trends and consistent low frequency variance drops substantially, to typically about 10 to 20%. Clearly the variance consistency test proposed here can pose a challenging test for the current models. We have plans to explore other types of variance consistency tests in our future work.

For the seven-model CMIP5 sample (Fig. 17 e), the percent of analyzed global area with attributable anthropogenic trends (including trends that are detectable but greater than simulated) is above 75% for 1901-2010 trends, for four of the seven models, with the remaining three models having lower percent area (45-75%). All seven models end up in the range of 40-70% for this metric for the latest starting date analyzed (1991). The metric that tests for both attributable anthropogenic trend and consistent low-frequency variance (Fig. 17 f), indicates that the seven models have a range of percent area of 17-35% for the 1901-2010 trends, but this range decreases to about 10% or less for the 1991-2010 trends.

7. Supplemental material and further sensitivity studies

The analysis presented in this study introduces a framework for trend analysis that has many possible applications and extensions. For surface temperature, there are many figures that are variations on the ones presented here, but were too numerous to include in this article. Therefore, we have created a web site based largely on this analysis, but which contains additional supplemental figures (http://www.gfdl.noaa.gov/surface-temperature-trends). For example, the web site contains plots for individual seasons that complement the annual-averaged analysis in this study. We show plots using alternative percentiles (97.5^{th} and 2.5^{th}) instead of 95^{th} and 5^{th}, and plots excluding certain low variability models from the analysis, etc. Additional regional plots like Figs. 7-9, including ones for individual seasons, are available, as well as maps

for different trend start dates. In addition, a number of plots based on analysis of individual CMIP3 or CMIP5 models, as opposed to multi-model ensemble means, are available.

8. Summary and Conclusions

The purpose of this analysis has been to introduce and apply a framework for assessing regional surface temperature trends using both the CMIP3 and CMIP5 models and using a multi-model sampling approach. We examined the behavior of the various control runs for the CMIP3 and CMIP5 models, and used the control run variability to help assess whether observed trends were unusual or not compared with the models' internally generated variability. We also used the control run variability to help assess whether observed trends were consistent with trends from the historical (20C3M) simulations—either All-Forcing runs or Natural-Forcing-Only runs. In cases for the CMIP5 models where trends were demonstrated to be inconsistent with Natural-Forcing-Only runs, but consistent with the All-Forcing runs, we conclude that an attributable anthropogenic component is present in the observed trend. For cases, such as the CMIP3 model assessments, where Natural-Forcing-Only runs are generally not available, we tested for detectable trends (compared to internal climate variability) and for consistency between observed and All-Forcing historical (20C3M) run trends.

In the separate CMIP3 and CMIP5 analyses, we generally attempt to give different models equal weight, even when a modeling center provides fewer ensemble members or shorter control runs. Tests are applied at global and regional scales, as well as at individual grid points on the observed data grid where there is sufficient data coverage over the period of the trend. Results are summarized using classification maps and global percent area statistics. Our analysis contains a substantial assessment of the variability in the models, including control run time series for visual inspection, standard deviation maps of low-pass filtered data, spectral analysis, and a low-frequency variance consistency test that is applied to individual models.

One of the most important results from the assessment is the identification of regions—and even grid points--where an anthropogenic warming signal is detectable in the observed temperature records. For trends over the period 1901-2010, a large fraction (about 80%) of the global area (with sufficient data coverage over time) has a detectable anthropogenic warming signal. Regions where the observed warming seems to be most commonly underestimated by the models include parts of the southern Ocean, south Atlantic, the far eastern North Atlantic, and off the east coast of Asia. The main regions without detectable warming signals include the mid- to high-latitude North Atlantic, the eastern U.S., and parts of the eastern and North Pacific. Moving forward in time, for the much shorter period (1981-2010) the observed warming trends over about 30% of the globe are assessed as having a detectable anthropogenic contribution. These regions include parts of the tropics, subtropics, and mid-latitudes (within about 40-45 degrees of the equator), and a narrow zonally oriented band near the Arctic Ocean. Areas without detectable trends (1981-2010) include much of the eastern Pacific--which is a region influenced by strong interannual variability associated with ENSO--and many extratropical regions poleward of about 40^{o}N and 40^{o}S. The CMIP3 models and the larger sample (23) of CMIP5 models yield results similar to those described above, although for these samples we assess only the consistency of trends, and not whether they contain an attributable anthropogenic component (due to the lack of Natural-Forcing runs with which to do such an assessment).

The reduced global area with detectable anthropogenic trends as one examines later start dates for trends in the record (all trends ending in 2010) illustrates the advantages of long records for trend detection in the context of this model-based assessment. In general, the shorter the epoch, the larger is the potential contribution of internal variability to the trend, leading to a greater spread (uncertainty) for sampled trends.

There are numerous examples of modeled trends or variability that are inconsistent with observations in our study. As has been noted in a previous paper using a similar methodology (Knutson et al. 2006), disagreement between modeled and observed trends in this type of analysis can occur due to shortcomings of models (internal variability simulation; response to forcing), shortcomings of the specified historical forcings, or problems with the observed data. A certain fraction of area should be expected to have inconsistent results due to chance alone (see Knutson et al. 1999 for further discussion of global significance testing in this context). As a further example, Wu and Karoly (2007) and Wu (2010) have noted that disagreement between simulated and observed regional surface temperature trends can result from shortcomings of models in simulating the observed warming associated with the changes of the leading climate variability modes (such as the Arctic Oscillation).

Concerning observational uncertainty, the HadCRUT4 data set (Morice et al. 2012) contains 100 ensemble members that attempt to characterize the uncertainties in the observations. We have performed some preliminary tests using these ensembles to assess the spread of observed trend estimates. These tests thus far indicate that even at the regional scale, the spread in trend estimates due to observational uncertainties, as contained in the ensembles, is generally much smaller than the spread in model simulated trends due to the internal variability and differences in forced responses in the historical runs (e.g., Figs. 7-9). However, in some regions (e.g., Mexico), the uncertainty in the observations plays an important role in the assessment of detectable anthropogenic contributions to trends.

We have attempted to at least partially address the issue of model uncertainties in the simulation of internal climate variability and in the response to historical forcing by using multi-model ensembles and by assessing consistency of both trends and low-frequency variability. When we apply a two-dimensional screening test (assessing simultaneously the consistency of the trend and low-frequency variability) we find that most models tend to be challenged to be consistent on both tests. Overall, our variance consistency tests suggest that while the CMIP3 and CMIP5 models provide a plausible representation of internal climate variability, there is considerable scope for improvement in the model simulations of internal climate variability, apart from their simulation of trends and variability in response to various forcing agents. From a different perspective, Shin and Sardeshmukh (2011) have concluded that the CMIP3 models do not simulate historical trends of temperature and precipitation as realistically as do atmospheric models forced by observed trends in tropical SSTs—a problem they attribute to model errors as opposed to climate noise (internal variability).

The CMIP3 and CMIP5 simulations used here represent "ensembles of opportunity" which cannot necessarily be expected to represent the true structural uncertainty in the results, due to shortcomings/uncertainties in the models and climate forcings. The procedures in our paper assume that the intrinsic internal variability of climate has not changed significantly since pre-industrial times, as we are using control run variability from pre-industrial control runs for our forced-run consistency tests. If anthropogenic forcing had actually *weakened* the intrinsic

variability in the real world, then our estimated uncertainty range around the All-Forcing model responses would be too wide -- making it overly difficult to conclude that observations were inconsistent with the All-Forcing runs. Similarly, if anthropogenic forcing had actually *strengthened* the intrinsic variability in the real world, then our estimated uncertainty range around the All-Forcing model responses would be too narrow -- making it too easy to conclude that the observations were inconsistent with the All-Forcing runs.

While the above uncertainty issues lack a final resolution, the methodology shown here can at least help to quantify the uncertainties associated with the climate change detection and attribution problem. The results show that when CMIP3 and CMIP5 historical runs are confronted with observed surface temperature trends, across a wide range of trend start dates, at various geographical locations around the globe, and even down to the grid point scale, a pervasive warming signal is found that is generally much more consistent with simulations that include anthropogenic forcing than with simulations that include either no forcing changes (control runs) or that include only natural forcing agents (solar, volcanic). Our conclusions about detectable anthropogenic contributions to the trends provide further support for the claim of a substantial human influence on climate, via anthropogenic forcing agents such as increased greenhouse gases. A future enhancement of our analysis would include an attempt to quantify the contributions of specific natural and anthropogenic forcing agents, or subsets of agents, in the CMIP5 All-Forcing and Natural-Forcing-Only historical runs. This would provide a more direct assessment of the relative influence of different forcing agents on the observed temperature trends at the regional scale.

Acknowledgments. We thank the Met Office Hadley Centre and the Climatic Research Unit, Univ. of East Anglia, for making the HadCRUT4 data set available to the research community. We thank the modeling groups participating in CMIP3 and CMIP5, and PCMDI for generously making the model output used in our report available to the community, and we thank three anonymous reviewers for their helpful comments on the manuscript.

References

Allen, M. R., and P. A. Stott, 2003: Estimating signal amplitudes in optimal fingerprinting. Part I: Theory. *Clim. Dyn.*, **21,** 477-491.

Allen, M. R., and S.F.B. Tett, 1999: Checking for model consistency in optimal fingerprinting. *Clim. Dyn.*, **15,** 419-434.

Delworth, T. L., and T. R. Knutson, 2000: Simulation of early 20th Century global warming. *Science,* **287**(5461), 2246-2250.

Annan, J. D. and J. C. Hargreaves, 2010: Reliability of the CMIP3 ensemble, *Geophys. Res. Lett.*, **37,** L02703, doi:10.1029/2009GL041994.

Hasselmann, K., 1997: Multi-pattern fingerprint method for detection and attribution of climate change. *Clim. Dyn.*, **13,** 601-612.

Hegerl, G. C., F. W. Zwiers, P. Braconnot, N. P. Gillett, Y. Luo, J. A. Marengo Orsini, N. Nicholls, J. E. Penner, and P. A. Stott, 2007: Understanding and attributing climate change. In *Climate Change 2007: The Physical Science Basis.* [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 996 pp.

Hegerl, G. C., et al. 2009: Good practice guidance paper on detection and attribution related to anthropogenic climate change. Available from IPCC: www.ipcc.ch/pdf/supporting-material/ipcc_good_practice_guidance_paper_anthropogenic.pdf

Hegerl, G.C., H. v. Storch, K. Hasselmann, B. D. Santer, U. Cubasch, and P. D. Jones, 1996: Detecting greenhouse gas induced climate change with an optimal fingerprint method. *J. Climate*, **9,** 2281-2306.

Karoly, D.J., and Q. Wu, 2005: Detection of regional surface temperature trends. *J. Clim.*, **18**, 4337–4343.

Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby, 2011: Reassessing biases and other uncertainties in sea-surface temperature observations measured in situ since 1850, part 2: biases and homogenization. *J. Geophys. Res.*, **116,** D14104, doi:10.1029/2010JD015220.

Knutson, T.R., T.L. Delworth, K.W. Dixon, and R.J. Stouffer, 1999: Model assessment of regional surface temperature trends (1949-1997). *J.Geophys. Res.*, **104**, 30981–30996.

Knutson, T.R., et al., 2006: Assessment of twentieth-century regional surface temperature trends using the GFDL CM2 coupled models. *J. Clim.*, **19**, 1624–1651.

Mahlstein, I., R. Knutti, S. Solomon, and R. W. Portmann, 2011: Early onset of significant local warming in low latitude countries. *Environ. Res. Lett.*, **6,** 034009, doi:10.1088/1748-9326/6/034009.

Meehl, G. A. et al., 2007: The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bull. Amer. Meteor. Soc.* **88,** 1383–1394.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observation estimates: The HadCRUT5 data set. *J. Geophys. Res.*, 117, D08101, doi:10.1029/2011JD017187.

Portmann, R. W., S. Solomon and G.C. Hegerl, 2009: Spatial and seasonal patterns in climate change, temperature, and precipitation across the United States. *Proc. Nat. Acad. Sci.*, www.pnas.org/cgi/doi/10.1073/pnas.0808533106

Rind, D., M. Chin, G. Feingold, D. Streets, R. A. Kahn, S. E. Schwartz, and H. Yu, 2009: Modeling the effects of aerosols on climate. In *Atmospheric Aerosol Properties and Climate Impacts, A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research.* [Mian Chin, Ralph A. Kahn, and Stephen E.

Schwartz (eds.)]. National Aeronautics and Space Administration, Washington, D.C., USA.

Sakaguchi, K. X. Zeng, and M. A. Brunke, 2012: Temporal- and Spatial-scale dependence of three CMIP3 climate models in simulating the surface temperature trend in the twentieth century. *J. Climate,* **25**, 2456-2470.

Schneider, T., and I.M. Held, 2001: Discriminants of twentieth-century changes in Earth surface temperatures. *J. Clim.*, **14**, 249–254.

Shin, S.-I., and P. D. Sardeshmuhk, 2011: Critical influence of the pattern of tropical ocean warming on remote climate trends. *Clim. Dyn*., **36,** 1577-1591.

Stenchikov, G., T. L. Delworth, V. Ramaswamy, R. J. Stouffer, A. Wittenberg, and F. Zeng, 2009: Volcanic signals in oceans. *J. Geophys. Res.*, **114**, D16104. doi: 10.1029/2008JD011673.

Stouffer, R. J., S. Manabe, and K. Y. Vinnikov, 1994: Model assessment of the role of natural variability in recent global warming. *Nature,* **367,** 634-636.

Stouffer R. J., Hegerl G. C. and Tett S. F. B. (2000): A comparison of Surface Air Temperature Variability in Three 1000-Year coupled Ocean-Atmosphere Model Integrations. *J. Climate*, **13,** 513-537.

Taylor, K.E., R.J. Stouffer, and G.A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93,** 485-498.

Vecchi, G. A., and A. T. Wittenberg, 2010: El Nino and our future climate: Where do we stand? *Wiley Interdisciplinary Reviews: Climate Change*, **1**, 260-270. doi: 10.1002/wcc.33.

Wittenberg, A. T., 2009: Are historical records sufficient to constrain ENSO simulations? *Geophys. Res. Lett.*, **36**, L12702. doi: 10.1029/2009GL038710.

Wu, Q., and D. J. Karoly (2007): Implications of changes in the atmospheric circulation on the detection of regional surface air temperature trends, *Geophys. Res. Lett.,* **34,** L08703, doi:10.1029/2006GL028502.

Wu, Q. (2010): Associations of diurnal temperature range change with the leading climate variability modes during the Northern Hemisphere wintertime and their implication on the detection of regional climate trends, *J. Geophys. Res.*, **115,** D19101, doi:10.1029/2010JD014026.

Yang, X., A. Rosati, S. Zhang, T. L. Delworth, R. G. Gudgel, R. Zhang, G. Vecchi, W. Anderson, Y.-S. Chang, T. DelSole, K. Dixon, R. Msadek, W. F. Stern, A. Wittenberg, and F. Zeng, 2013: A predictable AMO-like pattern in GFDL's fully-coupled ensemble initialization and decadal forecasting system. *J. Climate*, in press. doi: 10.1175/JCLI-D-12-00231.1.

Fig. 1

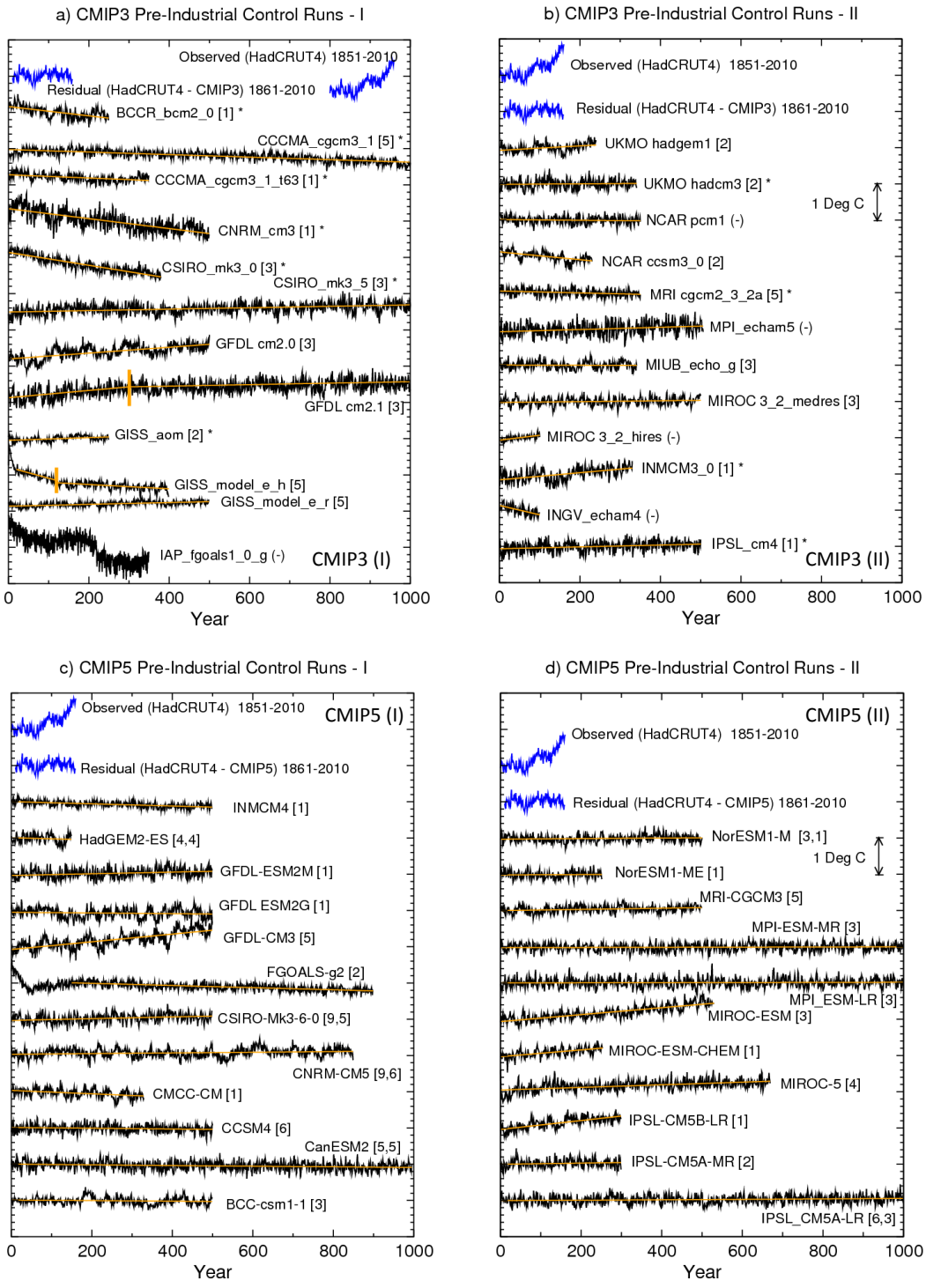
a) CMIP3 Pre-Industrial Control Runs - I
Observed (HadCRUT4) 1851-2010
Residual (HadCRUT4 - CMIP3) 1861-2010
BCCR_bcm2_0 [1] *
CCCMA_cgcm3_1 [5] *
CCCMA_cgcm3_1_t63 [1] *
CNRM_cm3 [1] *
CSIRO_mk3_0 [3] *
CSIRO_mk3_5 [3] *
GFDL cm2.0 [3]
GFDL cm2.1 [3]
GISS_aom [2] *
GISS_model_e_h [5]
GISS_model_e_r [5]
IAP_fgoals1_0_g (-)
CMIP3 (I)
0
200
400
600
800
1000
Year
b) CMIP3 Pre-Industrial Control Runs - II
Observed (HadCRUT4) 1851-2010
Residual (HadCRUT4 - CMIP3) 1861-2010
UKMO hadgem1 [2]
UKMO hadcm3 [2] *
1 Deg C
NCAR pcm1 (-)
NCAR ccsm3_0 [2]
MRI cgcm2_3_2a [5] *
MPI_echam5 (-)
MIUB_echo_g [3]
MIROC 3_2_medres [3]
MIROC 3_2_hires (-)
INMCM3_0 [1] *
INGV_echam4 (-)
IPSL_cm4 [1] *
CMIP3 (II)
0
200
400
600
800
1000
Year
c) CMIP5 Pre-Industrial Control Runs - I
Observed (HadCRUT4) 1851-2010
CMIP5 (I)
Residual (HadCRUT4 - CMIP5) 1861-2010
INMCM4 [1]
HadGEM2-ES [4,4]
GFDL-ESM2M [1]
GFDL ESM2G [1]
GFDL-CM3 [5]
FGOALS-g2 [2]
CSIRO-Mk3-6-0 [9,5]
CNRM-CM5 [9,6]
CMCC-CM [1]
CCSM4 [6]
CanESM2 [5,5]
BCC-csm1-1 [3]
0
200
400
600
800
1000
Year
d) CMIP5 Pre-Industrial Control Runs - II
CMIP5 (II)
Observed (HadCRUT4) 1851-2010
Residual (HadCRUT4 - CMIP5) 1861-2010
NorESM1-M [3,1]
1 Deg C
NorESM1-ME [1]
MRI-CGCM3 [5]
MPI-ESM-MR [3]
MPI_ESM-LR [3]
MIROC-ESM [3]
MIROC-ESM-CHEM [1]
MIROC-5 [4]
IPSL-CM5B-LR [1]
IPSL-CM5A-MR [2]
IPSL_CM5A-LR [6,3]
0
200
400
600
800
1000
Year

Fig. 1. Time series of global-mean annual-mean surface air temperature (2 m) anomalies from the CMIP3 (a, b) and CMIP5 (c, d) preindustrial control runs (black curves). Observed global mean surface temperature (HadCRUT4, combining SST and land surface air temperature anomalies) is also shown in blue on the diagrams for comparison. The blue curves labeled "Residual (HadCRUT4…" were created by subtracting the multi-model ensemble mean surface temperature (using masked SSTs and land surface air temperatures from the 20C3M All-Forcing historical runs for either CMIP3 or CMIP5) from the observed temperature. Thin orange straight lines (one or two segments) through the control run time series depict the long term linear drift and the period included in the analysis. The long term drift over the year range shown is calculated at each grid point and then subtracted from the model control run series before performing further analysis in our study. Short vertical orange segments denote two places where specific control runs were divided into two separate segments and the linear drift computed separately for each segment. In those cases, the residuals from the drift were formed and then combined back into a single series. The various curves in the figure have been displaced vertically by arbitrary constants for visual clarity. Curves labeled with a '*' denote CMIP3 models that did not include volcanic forcing in their historical runs. The number in brackets by each model name denotes how many All-Forcing ensemble members were available; when there are two numbers in brackets, the second refers to the number of Natural-Forcing ensemble members. Curves labeled with a '(-)' were excluded from the remainder of our analysis due to various issues such as discontinuities in time series, short record length, or unavailable sea surface temperature data in the CMIP3 archive. Vertical axis tic mark spacing is 0.2^{o}C.

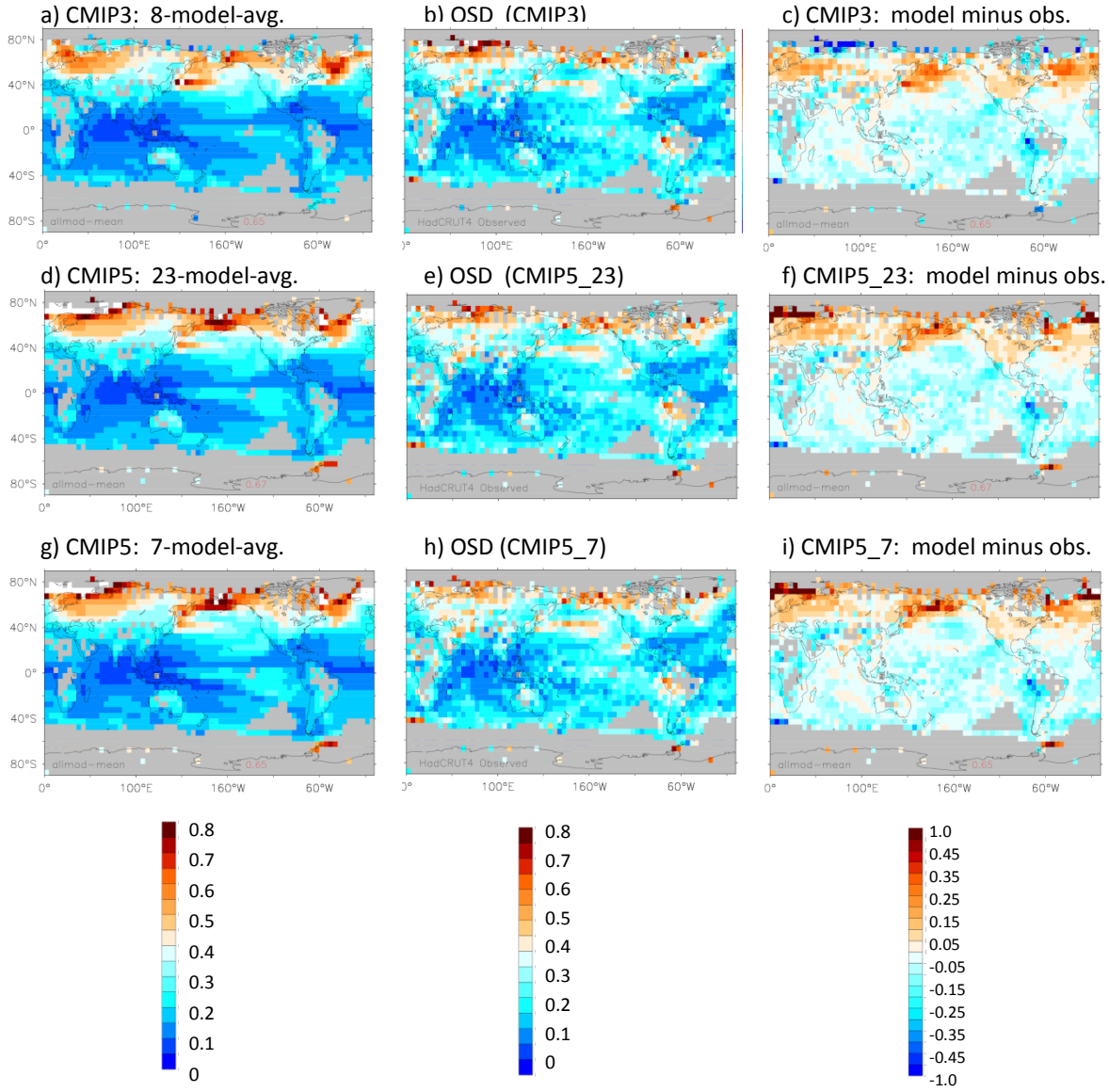

Fig. 2. Standard deviation (^{o}C) of low-pass (>10 yr) filtered internal variability of surface temperature derived from CMIP3 or CMIP5 pre-industrial control runs (a, d, g), an observed estimate (Obs. St. Dev.*; b, e, h), and the difference between the control runs and the observed estimate (c, f, i). The long-term linear drifts (time periods identified by the orange line segments in Fig. 1 a,b) were removed prior to computing the control run standard deviations. The model control run results are based on the mean standard deviation of a) eight CMIP3 models that have All-Forcing runs with volcanic forcing; b) all 23 CMIP5 models; and c) seven CMIP5 models that included at least one experiment with Natural Forcing only and extending to 2010. Note that the control runs on which the figure are based do not have episodic volcanic forcing and have been masked for observed missing data periods. Therefore, the observational estimate of internal variability (Obs. St. Dev.*) is derived from observations with adjustments for variance associated with various natural or anthropogenic forcing agents. See text for details of the adjustment.

Fig. 3

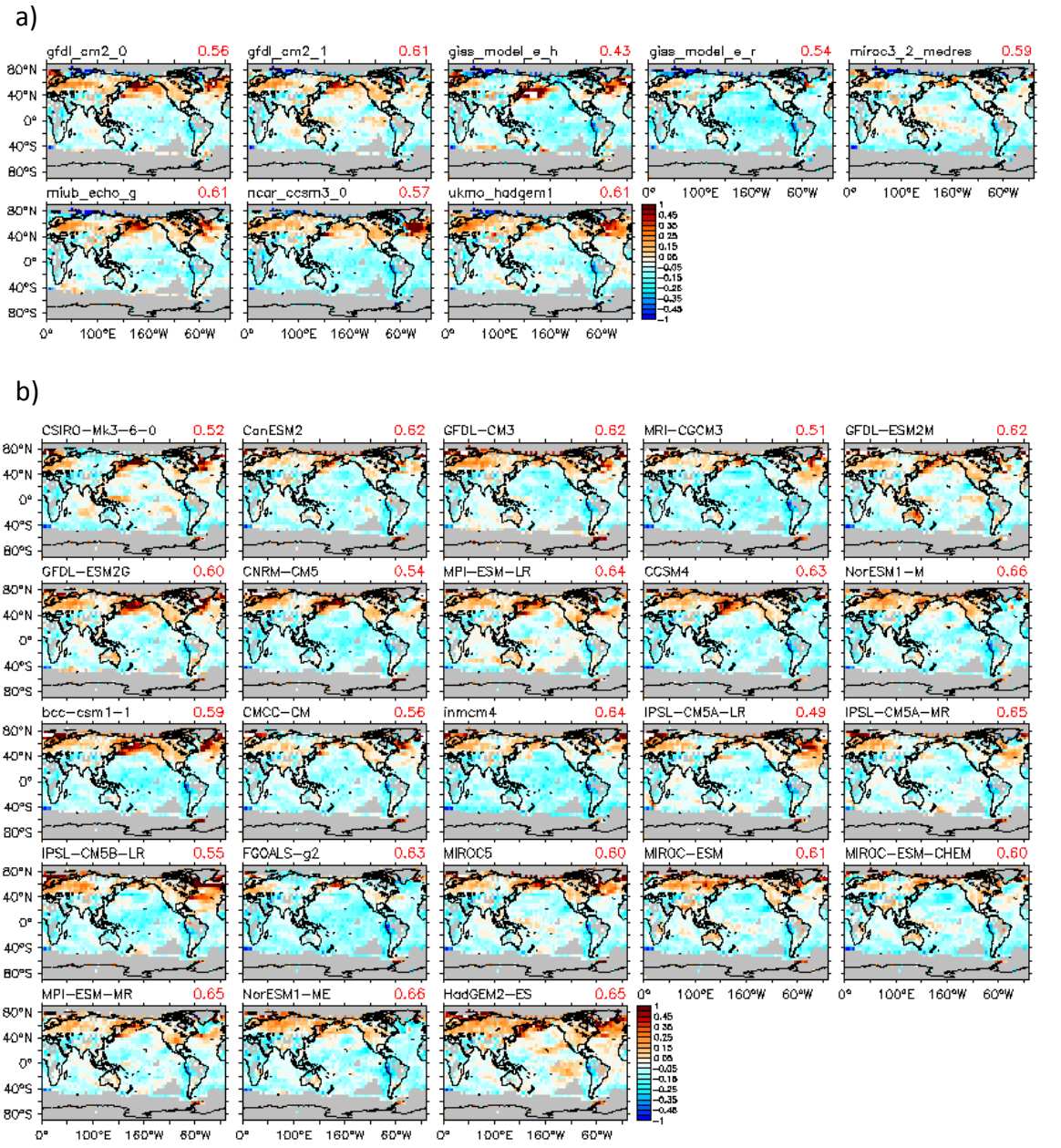

Fig. 3. As in Figure 2 (c, f, i) except for individual models in the (a) CMIP3 or (b) CMIP5 sets of models used in Fig. 2. The red number at upper right above each figure lists the spatial correlation of the model's low-pass filtered standard deviation field vs. the observational estimate (Obs. St. Dev.*) in Fig. 2. For the seven-model subset of CMIP5 models, the comparison is with the observations adjusted according to just those seven models and their respective All-Forcing and control runs. See text for further details.

Fig. 4

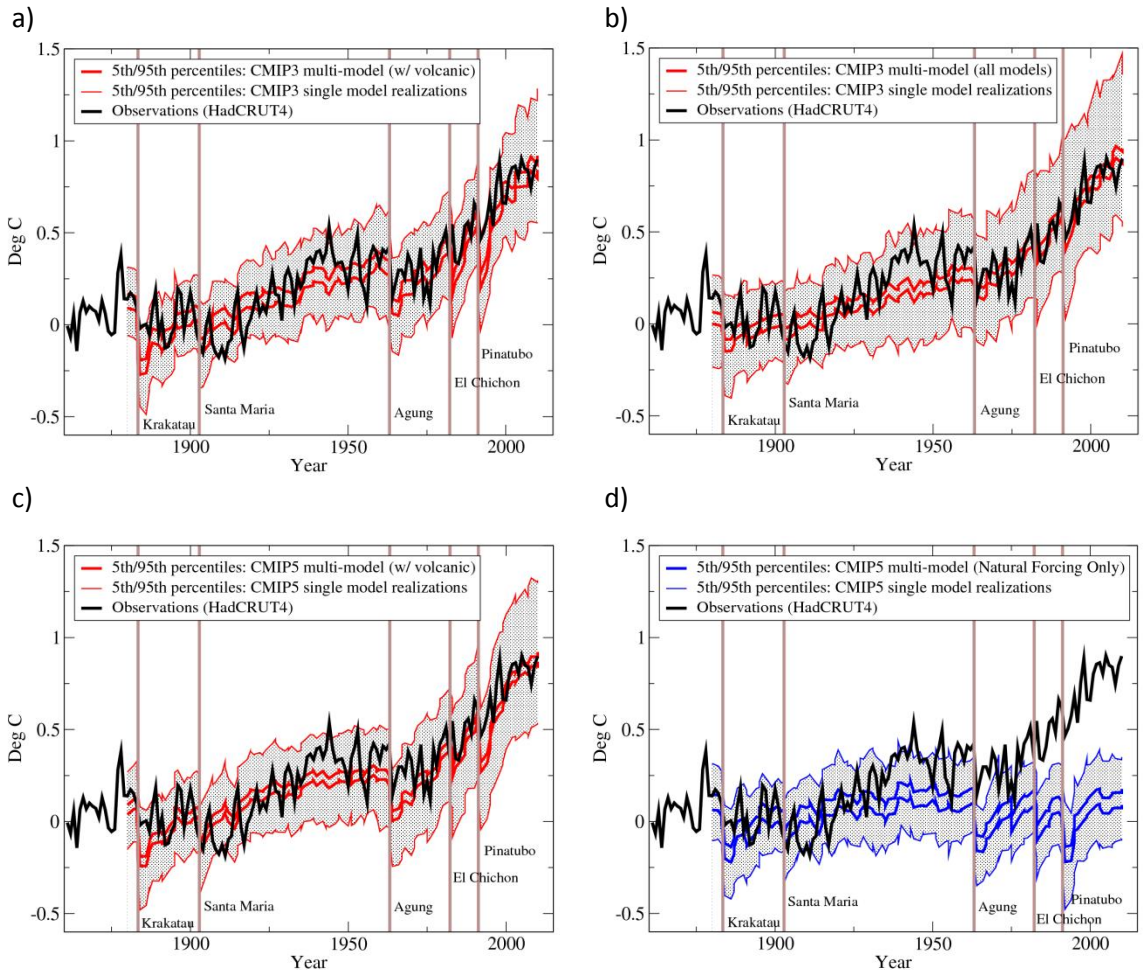

Fig. 4. Time series of global mean surface temperature anomalies (combined SST and land surface air temperature) from observations (HadCRUT4; black curves) in degrees Celsius. The red curves in a-c depict the 5th and 95th percentiles of annual mean anomalies for the multi-model mean (thick) or of single model realizations (thin lines, gray stippling) for the CMIP3 (a, b) or CMIP5 (c) 20C3M historical All-Forcing runs in degrees Celsius. The mean curve is not shown but lies approximately midway between the 5th and 95th percentiles. The series in (a) are from eight CMIP3 models run with volcanic forcing. The historical runs in (b) include 19 CMIP3 models with and without volcanic forcing (as identified in Fig. 1 (a,b). All of the 23 CMIP5 model runs included in the computations (c) incorporated volcanic forcing. In (d) the blue curves are based on seven CMIP5 models that had Natural-Forcing-Only runs extending through 2010. See text for description of how the confidence limits were computed. The time series have been re-centered so that the ensemble mean value, averaged for the years 1881-1920, is zero. Model data were masked with the observed spatially and temporally evolving missing data mask. The total number of individual experiments included in each panel was: a) 26; b) 51; c) 78; and d) 25.

Fig. 5

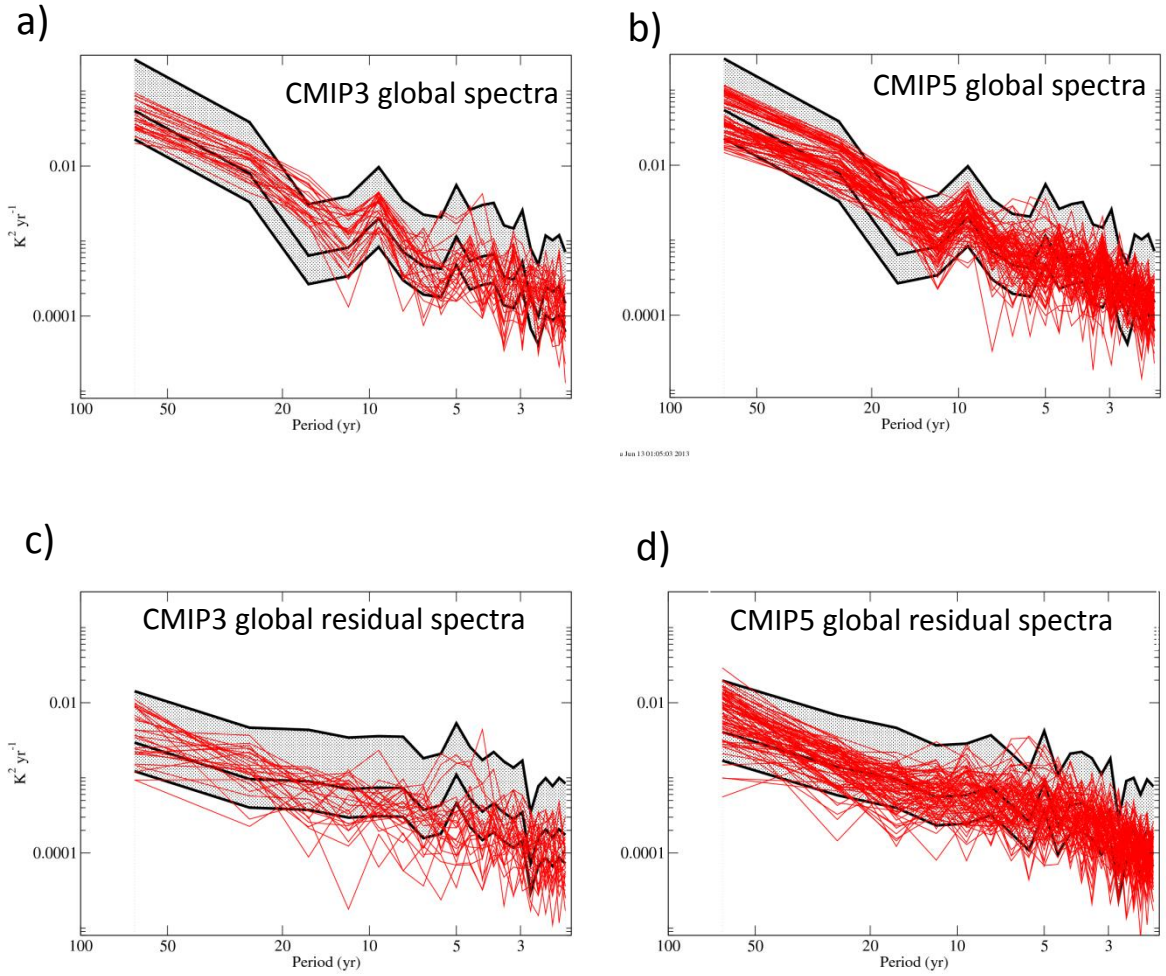

Fig. 5. Variance spectra as a function of frequency for observed global mean surface temperature (combined SST and land surface air temperature), in black with 90% confidence intervals shown by the shading, plotted against spectra for the individual (a) CMIP3 and (b) CMIP5 All-Forcing historical runs with Volcanic forcing (red) based on the time series in Fig. 4 (a,c). The spectra in (c) and (d) are based on residual observed or model historical run time series, where the multi-model ensemble surface temperature from the 20C3M All-Forcing (with volcanic) historical runs (CMIP3 or CMIP5) is subtracted from the observed and from each model run's global mean temperature series to form residual time series prior to computing the spectra (see text for details).

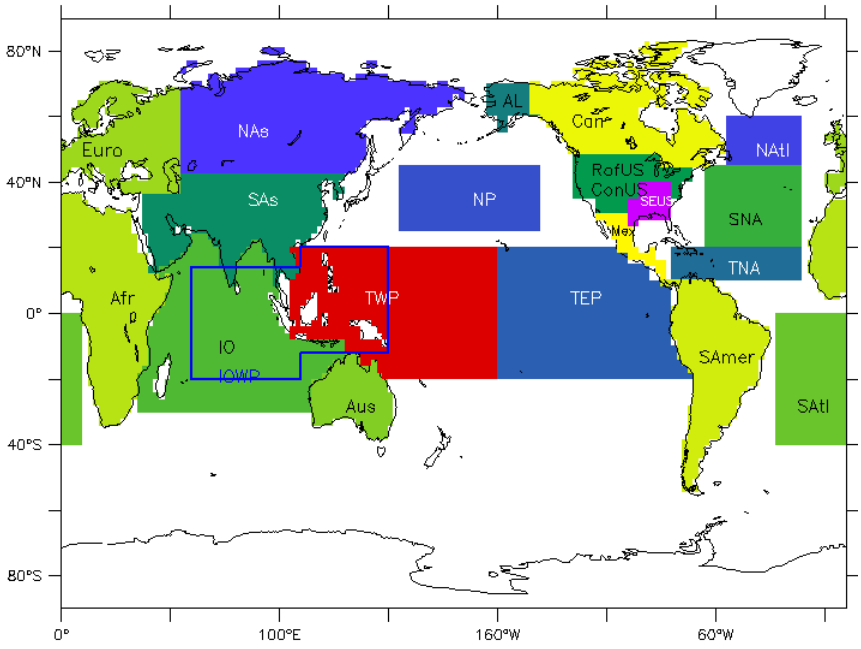

Fig. 6. Map illustrating averaging regions examined in Figs. 7-9. Regions abbreviations including: Euro = Europe; NAs = Northern Asia; SAs = Southern Asia; Afr = Africa; IO = Indian Ocean; Aus = Australia; TWP = Tropical western Pacific; TEP = Tropical eastern Pacific; IOWP = Tropical Indian Ocean/western Pacific warm pool; NP = North Pacific; AL = Alaska; SEUS = Southeastern United States; ConUS = Continental United States; RofUS = rest of continental United States, other than SEUS; SAmer = South America; Can = Canada; NAtl = North Atlantic; SNA = Subtropical North Atlantic; TNA = Tropical North Atlantic (Main Development Region); SAtl = South Atlantic.

Fig. 7

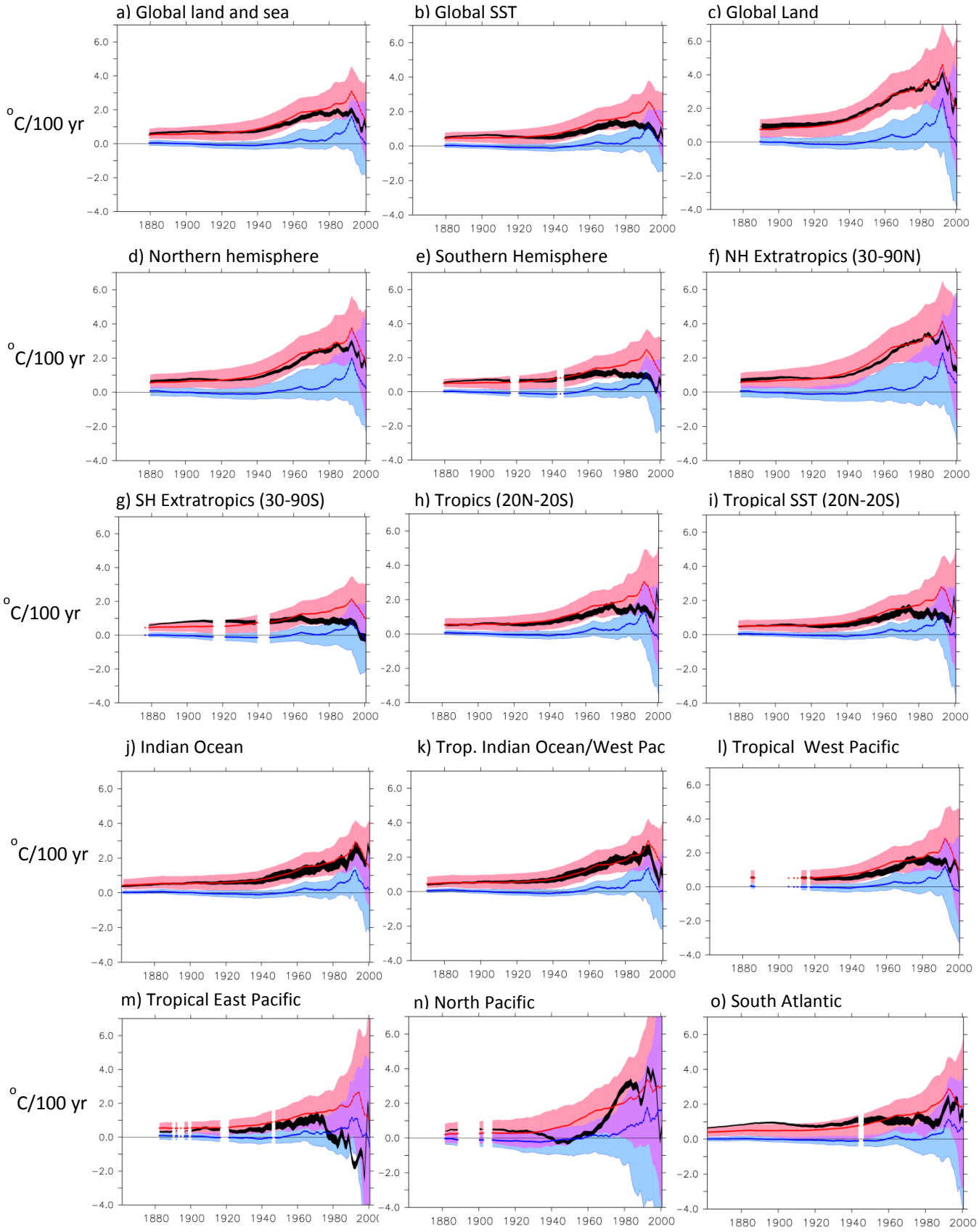
a) Global land and sea
b) Global SST
c) Global Land
d) Northern hemisphere
e) Southern Hemisphere
f) NH Extratropics (30-90N)
g) SH Extratropics (30-90S)
h) Tropics (20N-20S)
i) Tropical SST (20N-20S)
j) Indian Ocean
k) Trop. Indian Ocean/West Pac
l) Tropical West Pacific
m) Tropical East Pacific
n) North Pacific
o) South Atlantic
°C/100 yr
6.0
4.0
2.0
0.0
−2.0
−4.0
1880 1900 1920 1940 1960 1980 2000
Starting year of trend (All trends ending in 2010)

Fig. 7. Trends (^{o}C/100 yr) in area-averaged annual-mean surface temperature as a function of starting year, with all trends ending in 2010. The black curves are trends from observations (HadCRUT4), where observational uncertainty is depicted as a range showing the 5th to 95th percentile ranges of trends obtained using the 100-member HadCRUT4 ensemble. Blue curves are ensemble means for Natural-Forcing-Only runs using a subset of seven CMIP5 models that had Natural-Forcing runs to 2010. Red curves are ensemble means of the All-Forcing runs from the same seven CMIP5 models. See Fig. 6 for definitions of averaging regions. The different models are weighted equally for the multi-model ensemble means, regardless of the number of ensemble members they had. The pink shading shows the 5th to 95th percentile range of the distribution of trends obtained by combining random samples from each of the seven CMIP5 model control runs together with the corresponding model's ensemble-mean forced trend (All-Forcing runs) to create a total multi-model distribution of trends that reflects uncertainty in both the forced response and the influence of internal climate variability. The blue-shaded region shows the same, but for the Natural-Forcing-Only runs. Violet shading indicates where the pink- and blue-shaded regions overlap. Gaps in the curves indicate inadequate data coverage for a trend-to-2010 for those start years. Requirements include: 33% areal coverage to define an index time series point for a month, 40% of months available for a year to be non-missing, and 20% of all years available in each of five equal segments for a time series have adequate coverage for a trend. The seven-model CMIP5 subset used here and in subsequent assessment figures that incorporate Natural-Forcing runs include: CanESM2, CNRM-CM5, CSIRO-Mk3-6-0, FGOALS-g2, HadGEM2-ES, IPSL-CM5A-LR, and NorESM1-M.

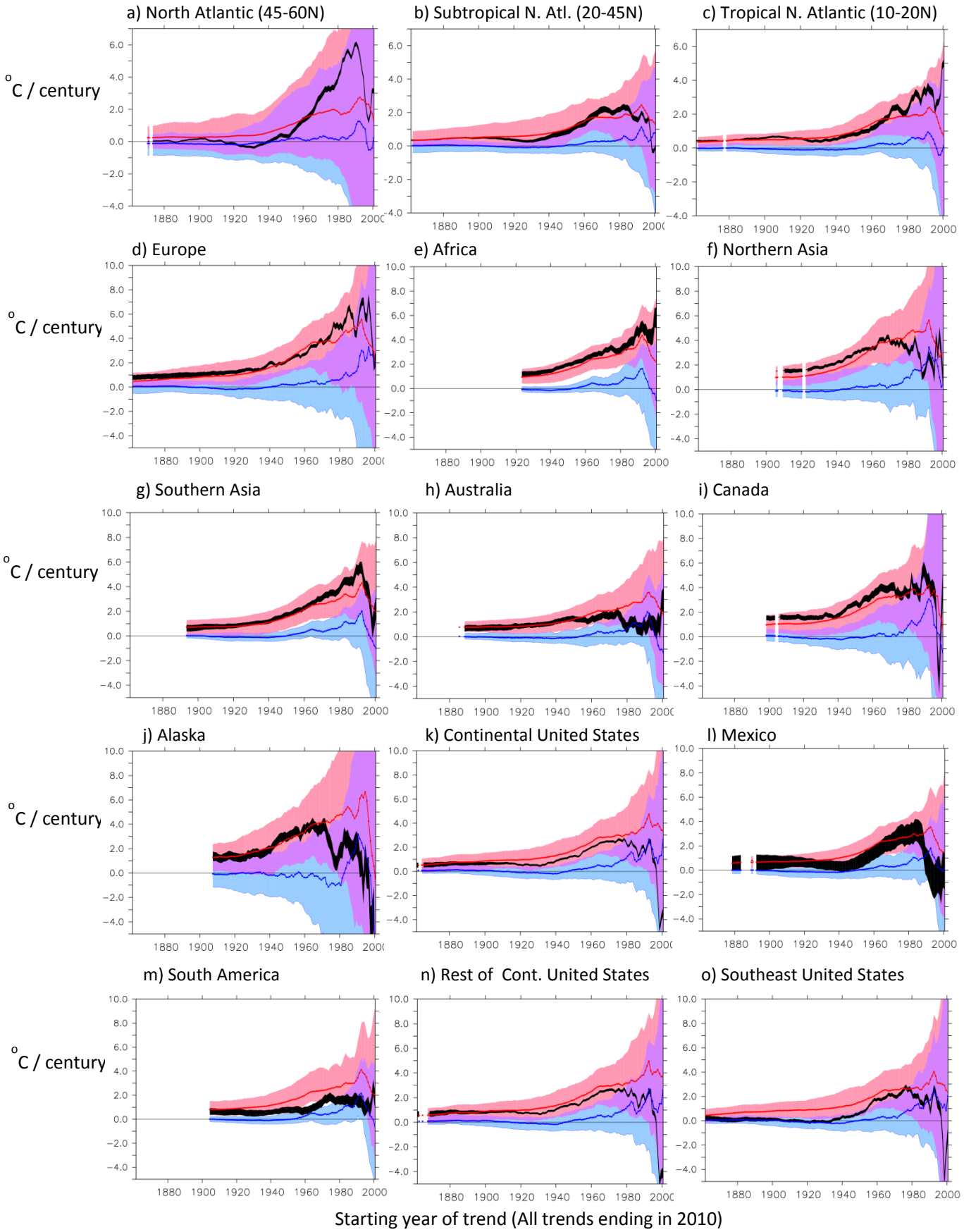

Fig. 8. As in Fig. 7, but for additional regions as labeled (see Fig. 6).

Fig. 9

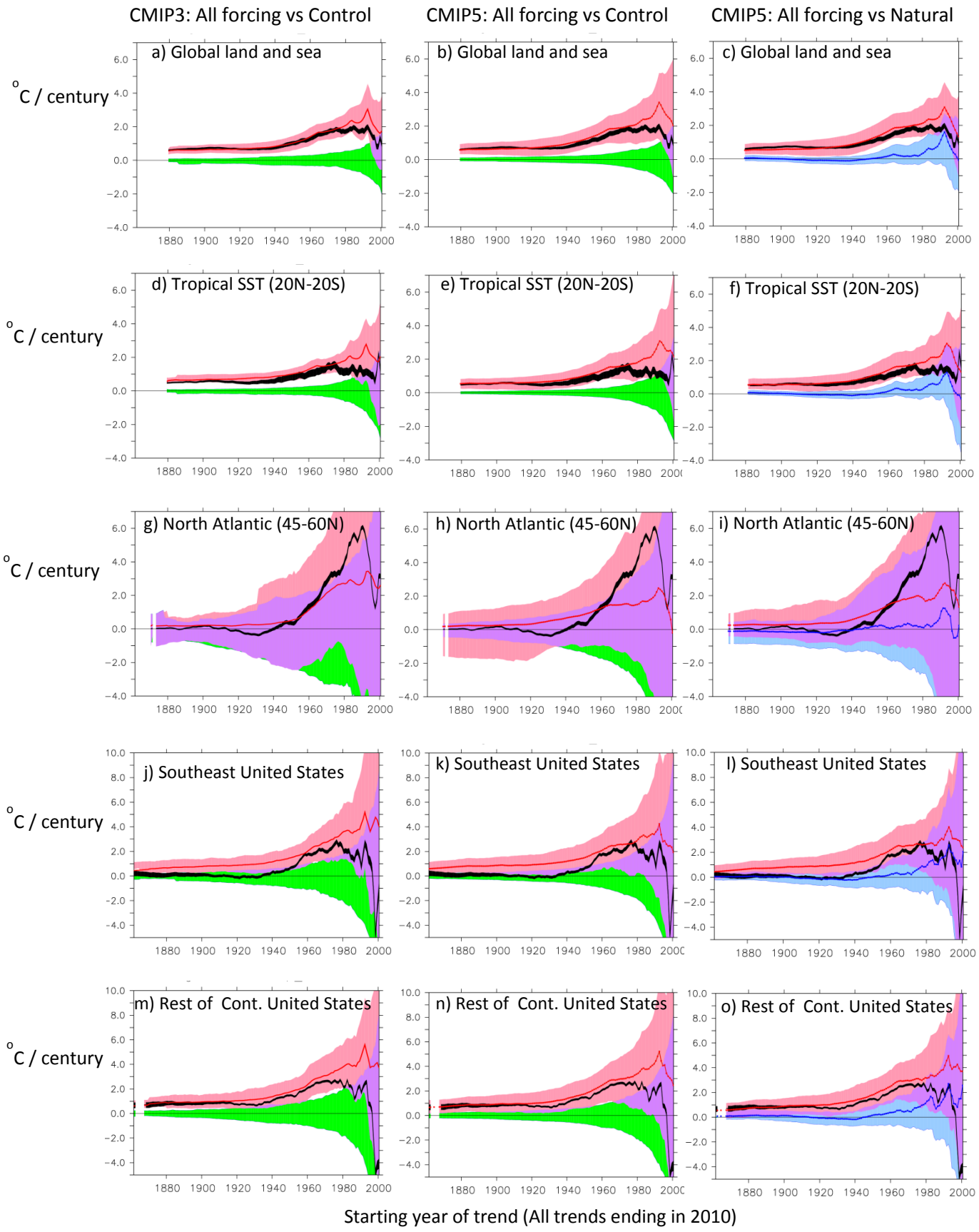
CMIP3: All forcing vs Control
CMIP5: All forcing vs Control
CMIP5: All forcing vs Natural
°C / century
a) Global land and sea
b) Global land and sea
c) Global land and sea
d) Tropical SST (20N-20S)
e) Tropical SST (20N-20S)
f) Tropical SST (20N-20S)
g) North Atlantic (45-60N)
h) North Atlantic (45-60N)
i) North Atlantic (45-60N)
j) Southeast United States
k) Southeast United States
l) Southeast United States
m) Rest of Cont. United States
n) Rest of Cont. United States
o) Rest of Cont. United States
Starting year of trend (All trends ending in 2010)

Fig. 9. As in Fig. 7, except the left column is based on All-Forcing runs from eight CMIP3 models that include volcanic forcing in their historical simulations, and the eight corresponding control runs (without volcanic forcing); the middle column is based on All-Forcing and control runs from all 23 CMIP5 models; and the right column is based on All-Forcing, Natural-Forcing-Only, and control runs from the same sets of CMIP5 models as used in Figs. 7 and 8 (see Fig. 7 caption).

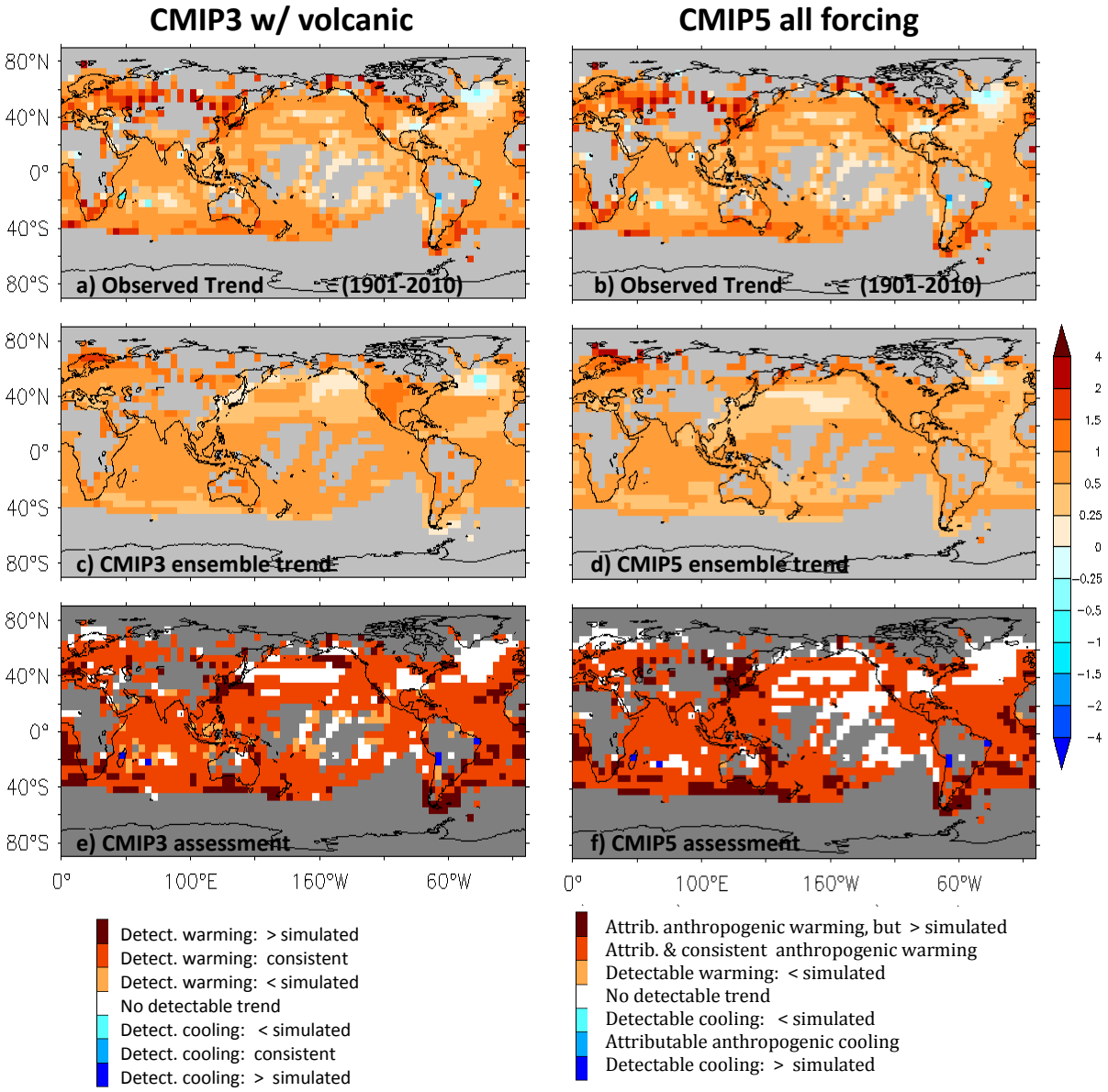

Fig. 10. Geographical distribution of surface temperature trends (1901-2010) in: (a,b) HadCRUT4 observations; (c) CMIP3 eight-model ensemble mean (All-Forcing, volcanic models); d) CMIP5 seven-model ensemble mean (All-Forcing, volcanic models). Unit: degrees C per 100 yr. In (e, f) the observed trend is assessed in terms of the multi-model ensemble mean trends and variability in the historical forcing and control runs (CMIP3 and CMIP5). The different colors in (e, f) depict different categories of assessment result; the categories are listed in the legends below panels e and f. Panel (e) compares observed trends with trends from eight CMIP3 All-Forcing models and their eight control runs. Panel (f) compares observed trends with trends from the CMIP5 seven-model subset, including All-Forcing, Natural-Forcing, and control runs.

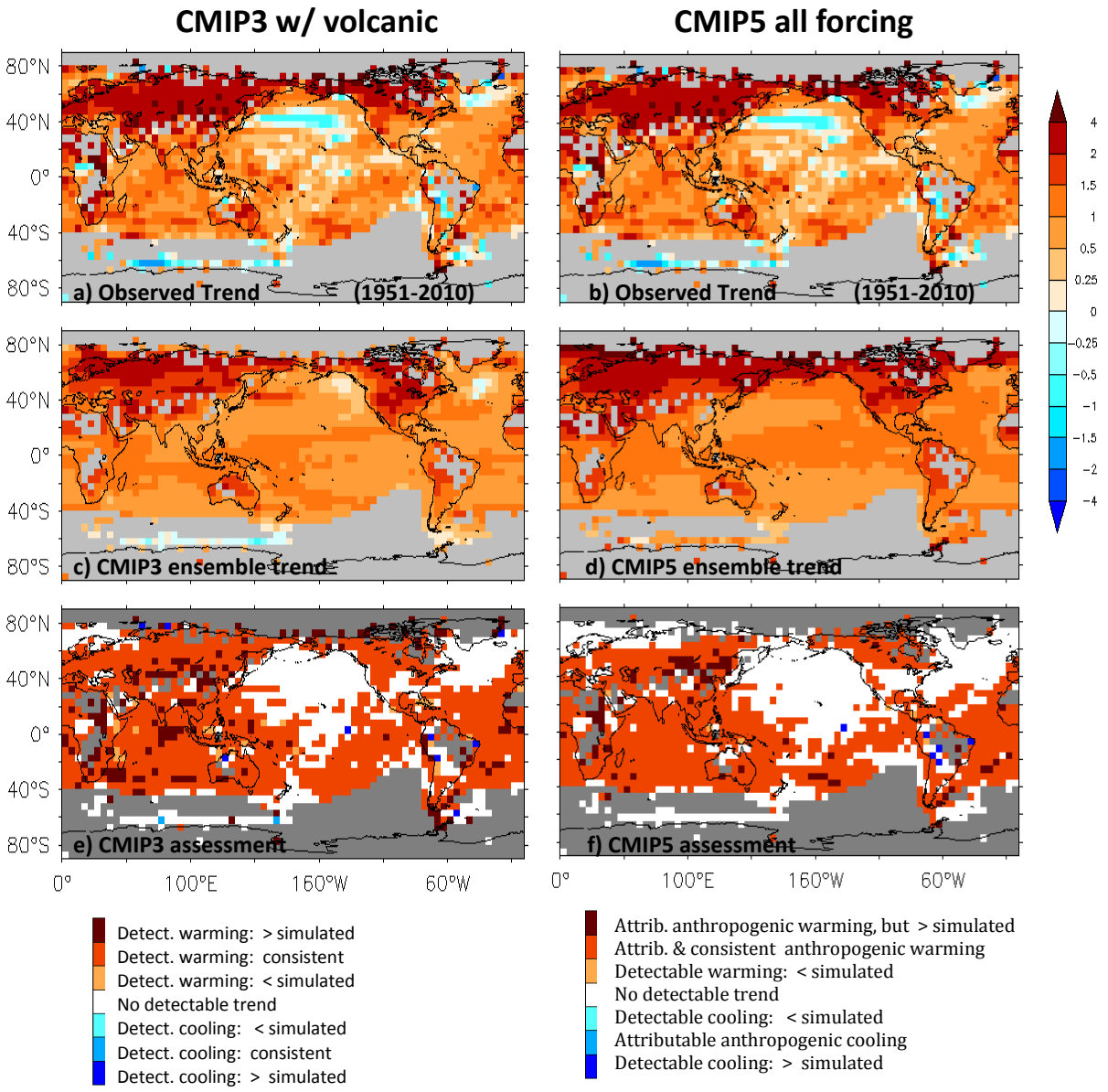

Fig. 11. Same as Fig. 10 but for trends from 1951 to 2010.

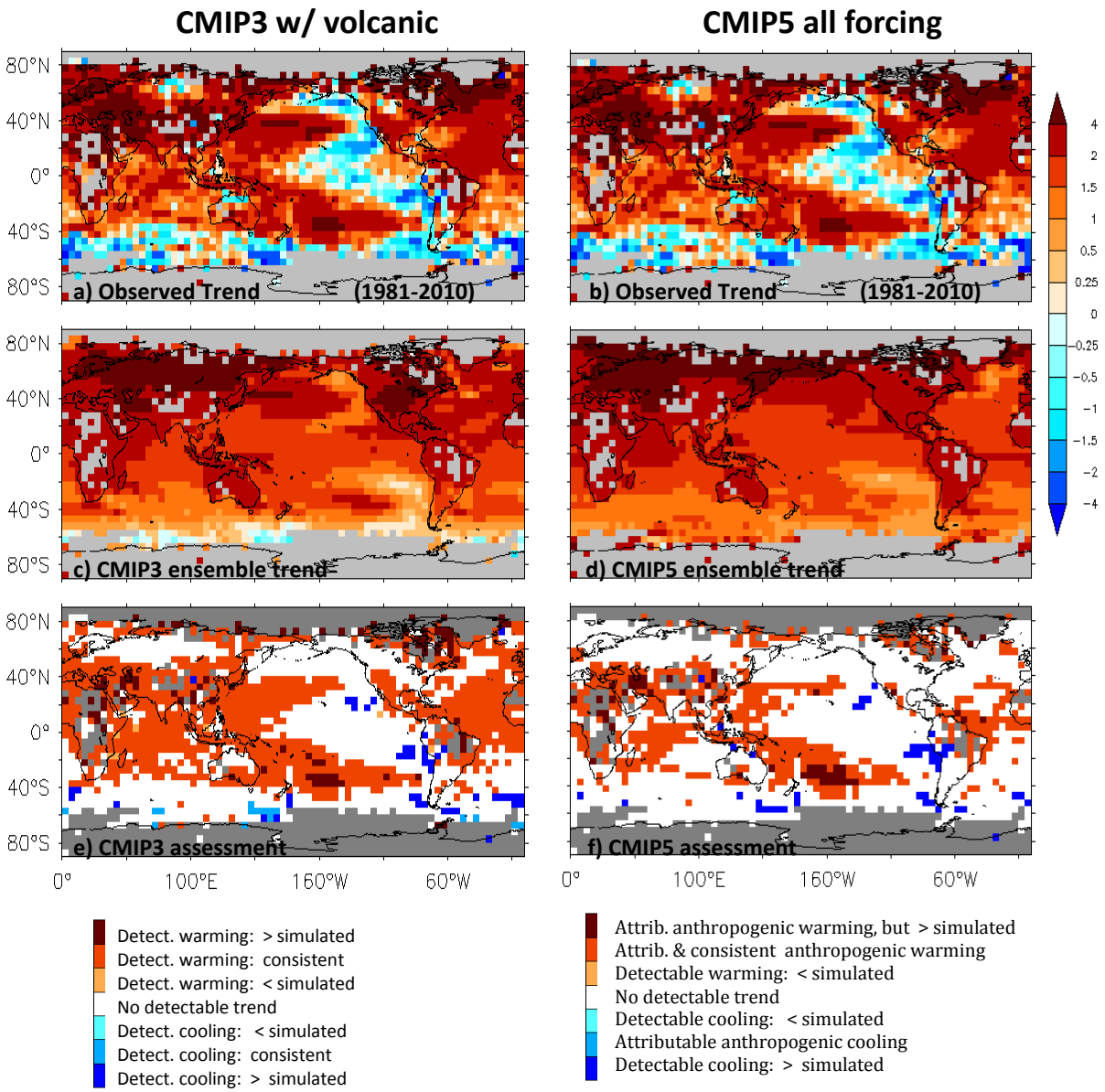

Fig. 12. Same as Fig. 10 but for trends from 1981 to 2010.

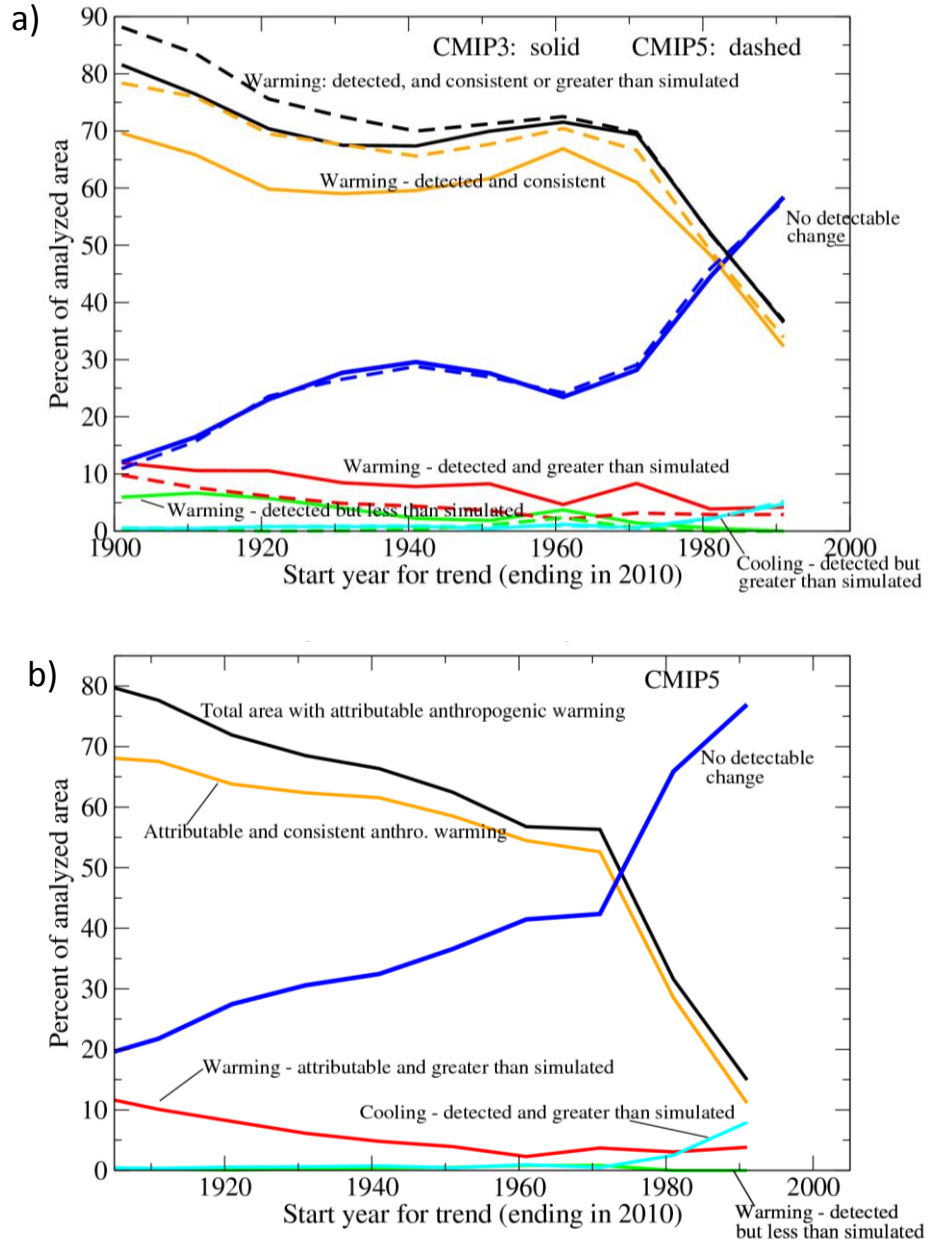

Fig. 13. Summary assessment of observed vs. model ensemble-mean trends-to-2010. The percent of global analyzed areas meeting certain criteria (see graph labels) are shown as a function of start year (all trends ending in 2010). a) Assessments of the eight CMIP3 (solid lines) vs. the 23 CMIP5 (dashed lines) multi-model ensemble means (historical 20C3M All-Forcing runs with volcanic forcing and associated control runs). b) Assessment of the CMIP5 multi-model ensemble means and control runs using the seven-model subset of CMIP5 models (with Natural-Forcing-Only runs extending to 2010), the All-Forcing runs from the same seven models, and their seven control runs. The black curves are the sum of the red and orange curves; the sum of black + cyan + green + blue = 100%.

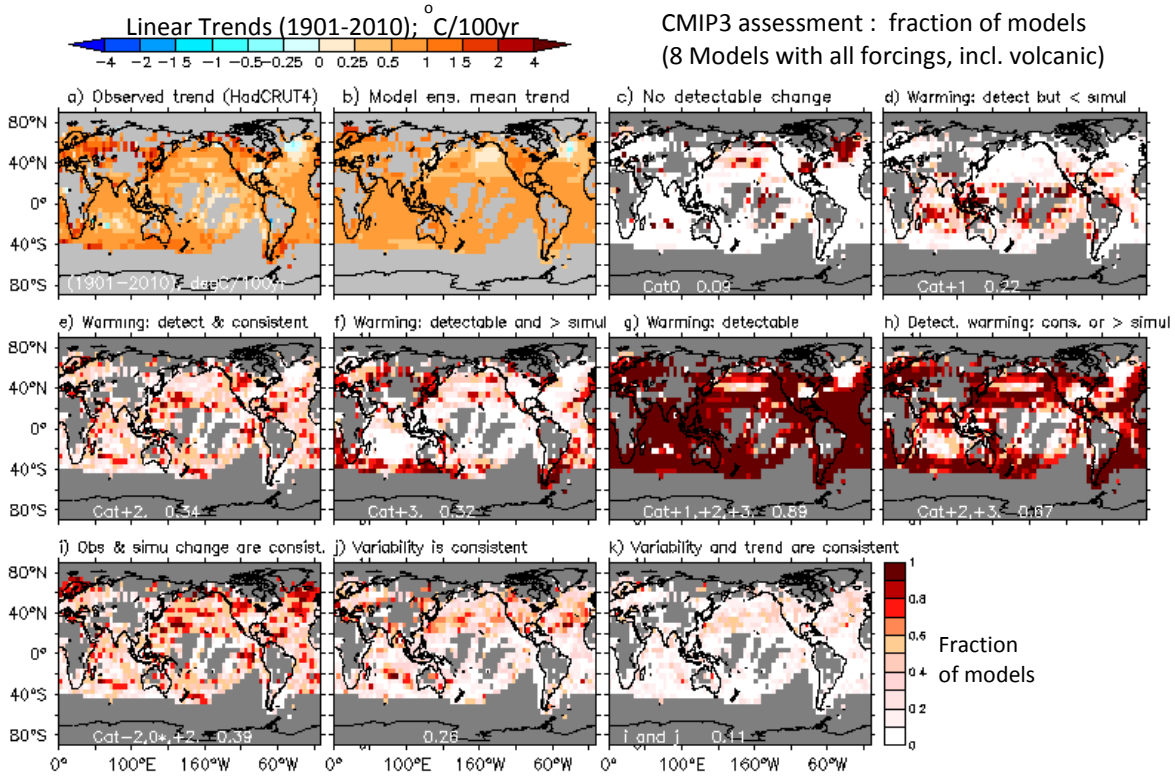

Fig. 14. Geographical distribution of: (a) HadCRUT4 observed and (b) CMIP3 multi-model (volcanic models) ensemble-mean surface temperature trends [^{o}C (100 yr)$^{-1}$, 1901-2010]. (c)-(k) The observed trend is assessed in terms of the eight individual CMIP3 models (trends and variability). The shaded field is the fraction of the eight individual CMIP3 models whose historical all-forcing runs meet the criteria listed above each panel. The criteria are as follows: c) no detectable change; d) warming that is detectable but significantly less than simulated in the all-forcing runs; e) warming that is detectable and consistent with the all-forcing runs; f) warming that is detectable but significantly greater than simulated in the all-forcing runs; g) warming that is detectable; h) warming that is detectable and either consistent with or greater than the simulated (all-forcing) runs; i) observed and simulated trends are consistent (though the observed trend may not be detectable); j) observed and simulated internal low-frequency variability are consistent; and k) conditions for (i) and (j) where both are satisfied (i.e., the simulated variability and trend are both consistent with observations). The white numbers at the bottom of maps in (c)-(k) indicate the area-weighted global average of the mapped fields.

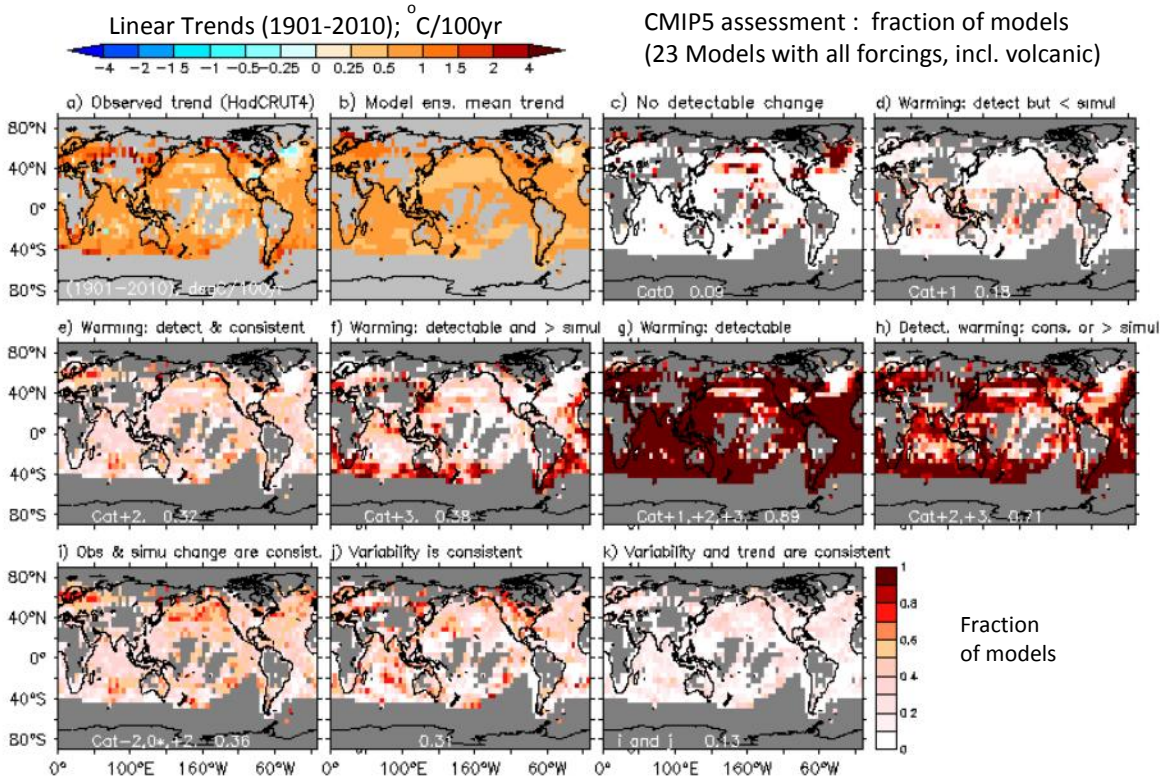

Figure 15. Same as Fig. 14, but for 23 CMIP5 models with volcanic forcing.

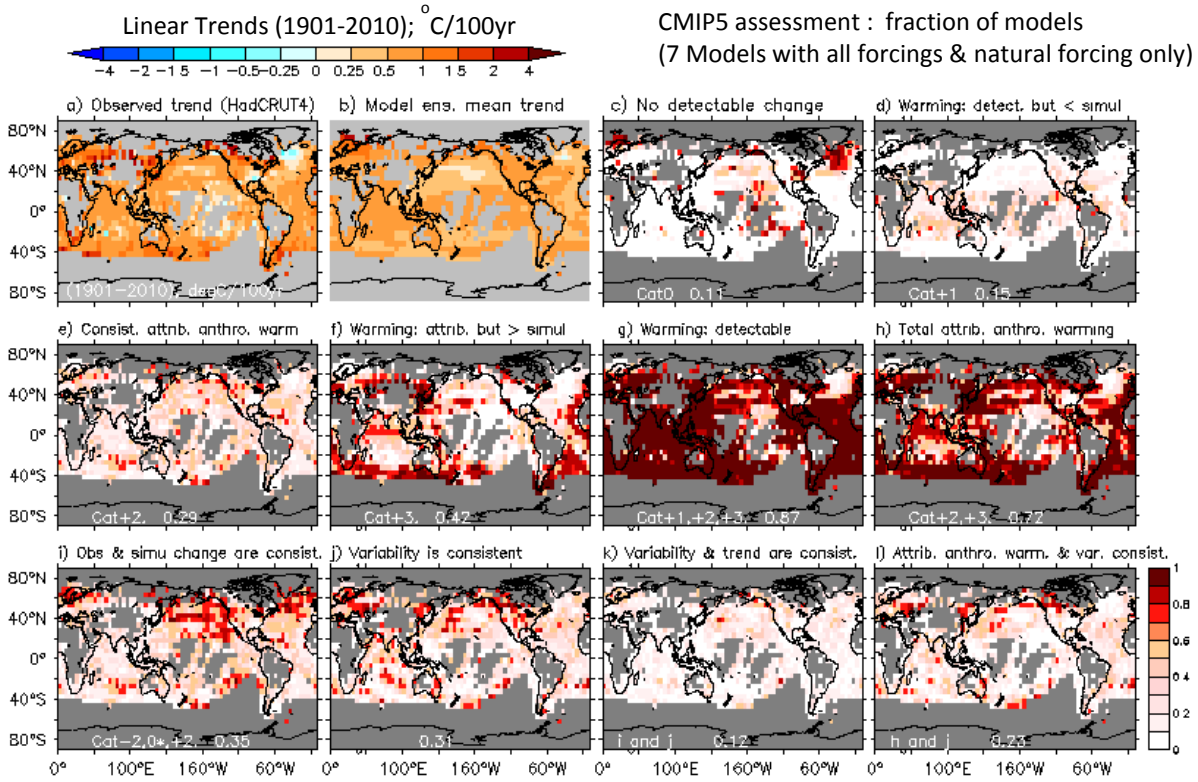

Fig. 16. Geographical distribution of: (a) HadCRUT4 observed or (b) CMIP5 multi-model ensemble-mean surface temperature trends (1901-2010) in degrees C per 100 yr. The observed trend is assessed in terms of trend and variability using the seven CMIP5 models that had available an All-Forcing ensemble and Natural-Forcing-Only runs extending to 2010. Panels (c-l) show the fraction of the seven individual CMIP5 models at each grid point whose All-Forcing, Natural-Forcing-Only, and control runs together meet the criteria listed above the panel. The criteria are: c) no detectable change; d) warming that is detectable (inconsistent with Natural-Forcing runs) but significantly less than simulated in the All-Forcing runs; e) attributable anthropogenic warming that is detectable (inconsistent with Natural-Forcing Only runs) and consistent with the All-Forcing runs; f) attributable anthropogenic warming that is significantly greater than simulated in the All-Forcing runs; g) warming that is detectable; h) total attributable to anthropogenic warming (i.e., sum of (e) and (f); i) observed and simulated trends are consistent (though the observed trend may not be detectable); j) observed and simulated internal low-frequency variability are consistent; k) conditions for (i) and (j) are both satisfied (i.e., the simulated variability and trend are both consistent with observations; and l) conditions for (h) and (j) are both satisfied (i.e., there is attributable anthropogenic warming and low-frequency variance is consistent with observations).

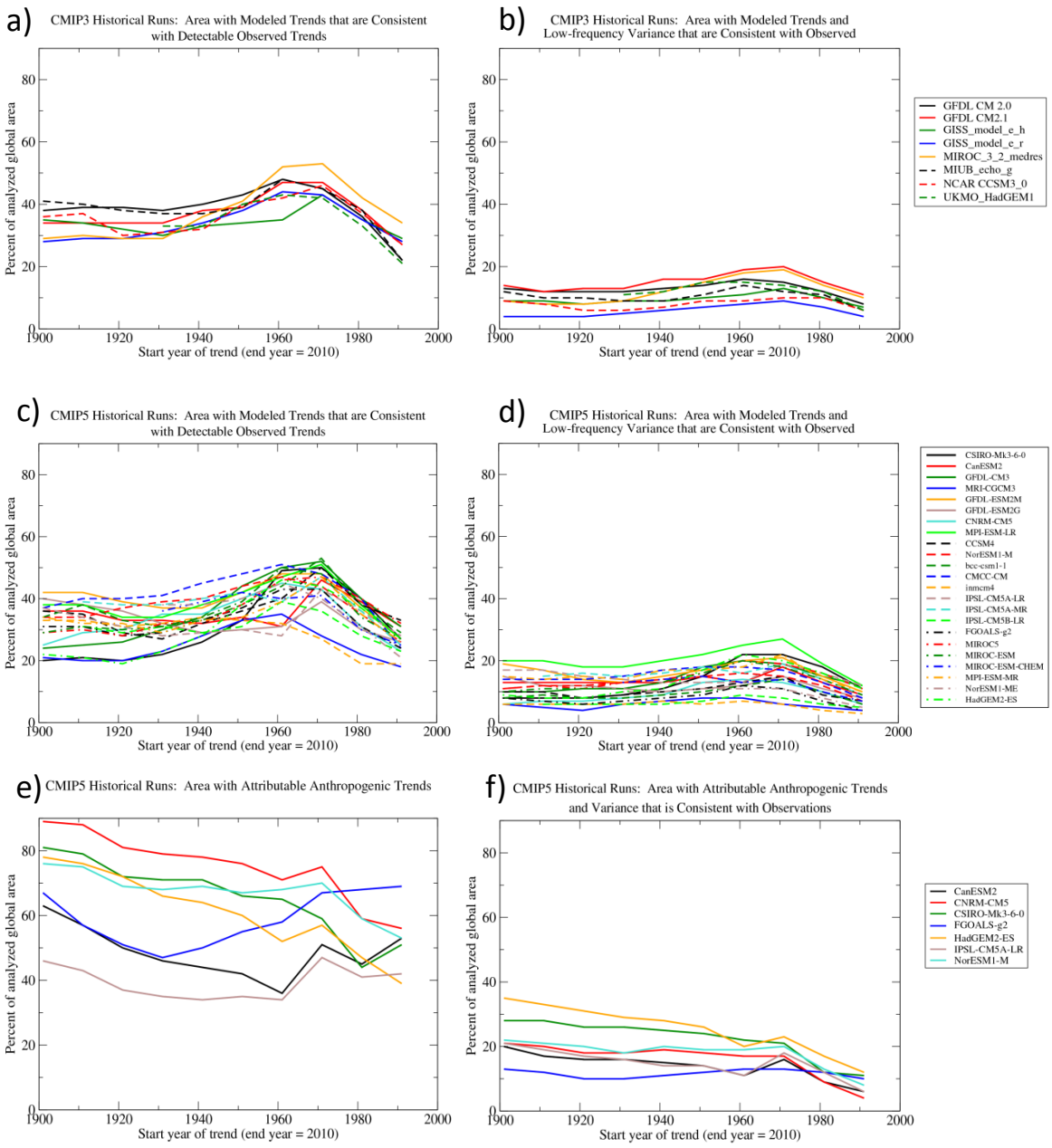

Fig. 17. Individual CMIP3 (a, b) and CMIP5 (c-f) models are assessed for (a), (c) consistency with detectable observed surface temperature trends-to-2010 (a-d), for attributable anthropogenic trends (e, f), and for consistency of both simulated trend and internal variability with observed estimates (b, d, f). Trend results are shown for start years from 1901 to 1991 (all trends ending in 2010). Plotted is the percent of analyzed global area where each individual model's (see legends) multi-realization ensemble mean forced trend and internal variability meet the criteria listed above the panel. The trends are analyzed at each grid point where there is sufficient temporal data coverage for the trend in question (see text). Note that panels (e, f) include areas where the observed trend is detectable and either consistent with or greater than simulated, whereas panels (c, d) include only areas with observed trends that are detectable and consistent with simulations.